

When Alignment Breaks: On Inference Integrity in Text-to-Image Models

Becky Qiao Ling Lin, Fengjun Li, and Bo Luo

EECS/I2S The University of Kansas, Lawrence, KS, USA
{beckylin, fli, bluo}@ku.edu

Abstract. Text-to-image (TTI) models are increasingly deployed in automated content pipelines, yet it remains unclear whether they preserve inference integrity, and whether they avoid mode-collapse-like concentration that amplifies representational bias. We conduct a large-scale audit of five major TTI systems, generating over 25,000 images across individual, group, mitigation, and intersectional prompting conditions, using occupational gender representation as a proxy for bias and distributional stability. We introduce group-based prompting as an integrity stress test that exposes latent mode collapse masked by single-subject generation and induces large demographic shifts under benign reformulations. Moreover, alignment effects also fail to generalize under intersectional prompts. These findings show that current alignment mechanisms address bias through surface-level heuristics rather than enforcing stable distributional constraints, motivating integrity-focused and fairness/bias-aware auditing for secure TTI deployment.

Keywords: Text-to-image generation · bias · inference integrity

1 Introduction

Text-to-image (TTI) generative models, such as DALL·E and Midjourney, have transitioned from creative tools to infrastructural components in automated media pipelines, including digital marketing, education, and large-scale content generation [34]. With platforms such as Midjourney reporting user bases exceeding 20 million [15], these systems increasingly function as visual knowledge engines that shape perception and inform downstream decision-making. However, as TTI models are relied upon to generate representative content, a critical question arises: do they possess *Inference Integrity*? We define inference integrity as the security property that a generative model’s output distributions are (i) *invariant* to benign, task-preserving prompt variations, (ii) *faithfully reflect* the model’s internal knowledge state, and (iii) *consistent* with broadly accepted human knowledge and empirical reality. A model with inference integrity should not exhibit large distributional shifts when the semantic intent of the prompt is preserved, nor should its outputs diverge from well-established empirical reference distributions without justification.

While significant research has been conducted on large language model auditing [9,43], TTI systems present a distinct challenge. Current TTI models exhibit

a “masked” vulnerability that resembles *mode collapse* [27,13], where the generator defaults to a narrow, high-probability set of outputs while suppressing the diverse “long tail” [1,31,22]. For example, prompts like “a software engineer” or “a housekeeper” may produce images representing any gender. When the TTI model repeatedly responds to such prompts, a model with inference integrity is expected to produce gender representations that reflect the empirical reality or its internal knowledge (training data), rather than collapsing to a single prototypical mode, and to do so consistently across repeated samplings and prompt variations that preserve the underlying occupational task family. However, TTI models often produce outputs that *reflect/amplify* traditionally dominant gender associations, frequently returning male- or female-dominated results. Such patterns suggest the violation of inference integrity and the presence of implicit biases. As proposed in a recent work on text LLMs [48], such mode collapse may be attributed to the *typicality bias* introduced by human tendency to prefer more “typical” (familiar or prototypical) outputs during Reinforcement Learning from Human Feedback (RLHF). This data-level bias can inadvertently amplify stereotypes, e.g., associating certain occupations primarily with one gender, and lead the model to pick the most expected gender presentation for a given profession. This phenomenon is a fairness deficit and a failure of alignment robustness.

We argue that traditional single-image auditing is insufficient to measure a model’s true distributional behavior. Inspired by the verbalized sampling theory in LLMs [48], we introduce group-based prompting as an implicit distribution sampling method for model auditing. By compelling the model to render multiple entities simultaneously (i.e., a distribution of outputs), group-based prompting induces broader sampling of the model’s output distribution and reveals behaviors masked by single-subject samples. The divergence between single-instance generation and group-based generation also indicates alignment brittleness.

Meanwhile, we also observed a critical vulnerability in current TTI systems: the potential decoupling of surface-level safety alignment from the model’s underlying latent distribution. While newer models have implemented safety guardrails (e.g., gender diversity in “CEOs”), they can be escaped by a benign prompt reformulation (e.g., adding an intersectional adjective: “Young CEOs”). When such prompts induce large shifts in the output distribution, the system lacks inference integrity and enables *distributional evasion attacks*: an adversary can bypass safety mechanisms not through injection or jailbreaking, but by exploiting legitimate semantic perturbations to toggle the model’s alignment state. We therefore argue that a generative model with inference integrity must exhibit distributional stability under policy-compliant, task-preserving prompt reformulations.

We aim to answer the following research questions: **RQ1** How does prompt structure (individual vs. group) affect the distributional stability of demographic outputs across occupations? **RQ2** How do different TTI models and model versions within model families vary in their demographic output distributions when given identical prompts? **RQ3** To what extent do model output distributions diverge from empirical baselines (labor data and search imagery), and does this divergence vary with prompt structure? **RQ4** To what extent can prompt phrasing

influence or stabilize demographic output distributions, and how robust are these effects to benign paraphrasing? **RQ5** Does gender bias persist or change under intersectional prompts (e.g., Black/Asian/Young CEOs), and do such prompts expose degradation in model alignment?

To answer these questions, we systematically evaluate gender representation across five widely used TTI models with a focus on *verbalized sampling using prompt phrasing*. We also compare model outputs to two real-world benchmarks: (1) image results from search engines such as Google and Bing, which reflect dominant societal visual narratives (potentially the training data or internal knowledge of the TTI models), and (2) gender distribution data from the U.S. Bureau of Labor Statistics (BLS). We primarily scope the large-scale audit to binary gender categories to establish a *robust baseline*. We further investigate other aspects of representation bias, such as evidence of built-in mitigation and prompt-based interventions. Moreover, to address the complexity of bias beyond a single axis and to explore the potential of prompt-based distributional evasion, we also conduct a targeted intersectional test to examine the stability of gender representation in specific contexts and the robustness of the safety guardrails.

In summary, this work makes the following contributions:

1. We introduce *inference integrity* as a reliability and security property for text-to-image (TTI) models to capture the consistency of output distributions.
2. We propose *group-based prompting* as a black-box integrity stress test that exposes distributional behavior masked by traditional single-instance audits.
3. We conduct a large-scale audit of occupational gender representation across multiple commercial TTI models, showing that prompt structure induces distributional instability and can reverse apparent demographic trends, exposing integrity failures under benign paraphrases.
4. We analyze model evolution and empirical alignment, demonstrating that newer models may reduce but do not eliminate bias, and that consistency with real-world statistics remains unstable.
5. We evaluate the robustness of built-in fairness guardrails through intersectional prompts, and demonstrate that state-of-the-art mitigation strategies are brittle and fail when occupational prompts are combined with demographic cues.

2 Background and Related Work

Bias in Generative Models. Research on bias in generative AI has primarily centered on text LLMs, e.g., on gender, racial, and occupational biases [9,14]. Fairness auditing frameworks were developed, e.g., stereotype association tests, debiasing algorithms, and benchmark datasets [43,40]. Bias amplification also emerges across synthetic training cycles even in contexts unrelated to fairness [44]. Besides, LLMs also encode cultural priors that often align with Western-centric values. Cultural prompting has been proposed as a mitigation [38]. While they provide valuable insights, their direct translation to multimodal systems remains limited. Unlike text, images exert stronger persuasive and emotional effects, making representational fairness in TTI systems especially consequential.

Table 1: Summary of representative papers on bias in TTI models. Top: referred publications; bottom: arXiv papers.

	Individual vs. group comparison	prompt-structure effects	formal fairness metric	baselines	intersectional reliability	models	proposed mitigation ²
[1]	×	×	proportional disparity	BLS	standard distribution	2	×
[47]	×	×	×	×	×	3	×
[41]	dual-subject only	×	Stereotype Score	BLS	qualitative/single-instance	1	FairCritic
[20]	×	×	diffusion bias+amplification	×	×	2	×
[18]	×	×	×	BLS	prompt enumeration	3	×
[46]	×	triplet prmpt.	disparity, pmt-img dependency	×	×	1	×
[7]	×	×	×	closed-set/human	open-set discovery	3	×
[40]	×	×	distrib. bias, hallucination, etc.	×	trend analysis	>100 ¹	×
[19]	×	×	bias scores, discrimination, etc.	×	multi-dimension classify	8	×
[45]	×	×	discussion only	concept only	survey	×	×
[42]	×	×	discussion only	concept only	survey	×	×
Ours	first comparative analysis	plural, order, count, etc.	absolute parity+JSD	BLS+Google+Bing	prompt-based stress test	6	structural (plurality) & explicit prompting

¹ besides the foundation models, [40] also studied artistic & style-transferred models.

² × indicates the study did not systematically evaluate an intervention strategy.

Fairness in Text-to-Image Models. Compared to LLMs, investigations of fairness in TTI are more recent but growing, e.g., surveys highlight multidimensional biases and emphasize the lack of unified evaluation frameworks [42]. Recent work on dual-subject bias found that TTI models reinforce gender stereotypes when asked to generate two subjects in an occupational relationship [41]. Internal audits reveal that diffusion processes can introduce/exacerbate disparities, with Stable Diffusion v2 showing stronger skew than DALL·E 2 [20]. Prompt-based analyses find that gender indicators influence image layouts and object choices, with neutral prompts defaulting to masculine representations [46].

Despite these advances, most prior research remains either model-specific, limited in cross-model comparisons, or disconnected from empirical social baselines. As summarized in Table 1, prior work differs in its treatment of prompt structure, fairness metrics, baselines, intersectional reliability, and the number of models considered. Our study builds on these foundations by systematically auditing five widely used TTI systems and situating their representational outputs relative to both labor statistics and search engine images. We implement verbalized sampling using group-based prompts, aiming to mitigate mode collapse and trigger diversified outputs. Besides the new methodology (verbalized sampling, cross-model audit, Internet image as baselines, formal metrics, diversified prompt structure, prompt-based mitigation), we also present our new findings discovered through these new methods.

Prompt Sensitivity and Mitigation. Prompt design has emerged as a key factor in shaping generative model outputs. In LLMs, even subtle changes in phrasing influence the salience of encoded biases [33,28]. Studies of TTI suggest similar dynamics: modifiers, ordering, and plurality in prompts systematically alter demographic representation [22]. Commercial providers have begun implementing interventions at the prompt level. For example, DALL·E employs reinforcement learning with human feedback (RLHF) to encourage more diverse generations [23]. However, they can also lead to overcorrection or inconsistencies. We extend this line of research by systematically quantifying how singular versus plural phrasing affects bias in TTI outputs and by evaluating prompt engineering as a practical mitigation across multiple model families.

Intersectionality and Skin-Tone Bias in Vision/TTI. Prior research shows that demographic attributes rarely operate in isolation. Intersectional analyses reveal compounding disparities across gender and race/skin tone in computer vision and generative models [18,4,16]. In TTI models, prompts such as “a nurse” or “a CEO” can implicitly trigger defaults not only for gender but also for phenotypic traits and cultural markers (e.g., skin tone, attire). Empirical studies using balanced face datasets (e.g., FairFace) and skin-tone benchmarks show that model outputs often overrepresent lighter skin tones in high-status roles and darker tones in care or service roles, even when gender is held constant [4,16]. In this work, we use gender as our primary large-scale axis of analysis and complement it with a targeted intersectional stress test (e.g., race and age).

Search and Media as a Social Baseline. It is shown in the literature that biases in LLMs were inherited/amplified from large-scale internet training data [9,14]. In this work, we also aim to shed light on the source of the biases in TTI generation. We employ search engine images as a “social baseline”, which may potentially represent the internal knowledge of the TTI models [12,25,35].

Summary of Differences from Prior Work: Compared to existing studies, we (1) introduce inference integrity as a distributional stability property, (2) systematically evaluate prompt-structure effects (individual vs. group), (3) conduct a cross-model audit across multiple commercial and open-source TTI systems, (4) benchmark against both labor statistics and web image distributions, and (5) examine robustness under intersectional prompting.

3 Methodology

3.1 Prompt Design and Occupation

We selected 15 occupations based on US BLS data: five female-dominated ($\geq 2/3$ women), five gender-balanced ($\sim 50\%$ women), and five male-dominated ($\geq 2/3$ men) careers. We use two prompt types: single-instance (e.g., “a chief executive”) and group (e.g., “a group of chief executives”) prompts.

Group prompts require the model to render multiple individuals within a single image, eliciting broader demographic variation than single-instance prompts. Analogous to multi-output prompting in LLMs (e.g., “generate 10 jokes”), this forces the model to instantiate multiple entities within a shared inference context. Therefore, group prompts serve as a modality-appropriate analogue of verbalized sampling, enabling us to assess whether demographic patterns remain stable under increased compositional demand, complementary to repeated single-instance generations. We treat singular $\rightarrow$ group as task-preserving at the occupation concept level (same role semantics), while intentionally varying multiplicity/composition to probe distributional stability.

To minimize prompt-induced bias [33,28], all prompts were kept neutral and simple. Models were used with default public settings, e.g., Midjourney was accessed via version 6.1 (2025/05) [21], and DALL-E 3 through the OpenAI API

with default configuration, without any manual editing, post-processing, style modifiers (e.g., ‘photo of’, ‘illustration of’), or seed control. All data were collected during May to November 2025 using publicly available interfaces. Since several evaluated systems are proprietary services, backend updates and regional deployment details are not fully observable. We therefore interpret the results as a snapshot audit of deployed systems under the documented collection conditions. Each prompt produced 100 individual and 100 group images, resulting in 15,000 images in total. Examples are shown in Appendix A Figure 4.

3.2 TTI Models and Baselines

We selected five widely used TTI models to ensure coverage across commercial, community-driven, and open-source ecosystems. DALL·E v2 and v3: state-of-the-art commercial systems with known bias interventions; Midjourney: a large, active user base and a proprietary pipeline; DeepAI (<https://deepai.org/machine-learning-model/text2img>): a simpler API-based system with fewer constraints; and Stable Diffusion 3.5 (Large) [8]: an open-source baseline with partial training data transparency. These models span different architectures (transformer-based vs. diffusion-based) and adoption contexts, allowing a systematic comparison across the current TTI landscape.

To enable longitudinal analysis across proprietary and open ecosystems, we adopted Stable Diffusion XL (SDXL) 1.0 [29] as a historical baseline for the open-source models. We used three representative occupations: Chief Executive (male-dominated), Registered Nurse (female-dominated), and Bartender (gender-balanced), to examine generational shifts in a managed service (DALL·E 2 vs. 3) and an open-weights architecture (SDXL vs. SD 3.5).

We compare TTI outputs to Google/Bing image search results and to U.S. BLS data, which serve as reference distributions for evaluating inference integrity rather than normative fairness targets. These baselines represent two distinct but plausible sources of ground truth: (i) empirical workforce statistics (BLS), and (ii) dominant visual narratives available online (search engines). We retrieved the images using the same neutral prompts as used for the TTI models and collected all images returned for each search query (typically 400 to 600 images per prompt). By comparing AI-generated images to both sources, we assess whether demographic representations remain stable across semantic variation prompts or whether they shift in ways that indicate inference instability.

3.3 Face Detection and Classification

We used RetinaFace [5] for face detection. RetinaFace was selected for its robustness to pose variation, occlusion, and image artifacts common in generated images. Detected faces were cropped and classified using two models: InsightFace [30,11,10] and FairFace [16]. We adopt an ensemble approach to balance accuracy and bias considerations: InsightFace provides strong performance across diverse conditions, while FairFace is trained on a demographically balanced dataset. For individual-prompt images, a gender label was accepted only

when both models agreed; disagreements and images without detected faces were manually reviewed. For group-prompt images, where manual labeling is impractical due to the number of individuals, we averaged predictions from both models to obtain aggregate gender statistics. For example, if a group image contains 10 detected faces, with InsightFace classifying 6 as female and FairFace classifying 4 as female, we treat the image as contributing an estimated female proportion of $(6/10 + 4/10)/2 = 0.50$. Images without detectable faces as shown in Appendix A Figure 5 accounted for fewer than 5% of cases and were manually classified; these were primarily cartoon-style renderings, cropped faces, or back views.

To validate labeling accuracy, we conducted manual annotation on a subset of individual images sampled across all models. The annotation was performed independently by one author and two volunteer annotators. Disagreements were resolved through discussion. We compared manual labels with the ensemble predictions, yielding 96%–100% agreement. Most disagreements occurred in visually ambiguous cases, particularly in images generated by DALL·E 2, which more frequently exhibits distorted facial features (see Appendix Figure 6).

3.4 Training Data and Transparency

Most commercial TTI systems do not disclose their training data. Available documentation suggests reliance on large-scale web-scraped image–text corpora [6,32]. Open-weight models such as Stable Diffusion are known to use LAION-5B, which reflects biases present in internet imagery [32,2]. Because training data is largely opaque, we compare model outputs to Google and Bing image search results as a coarse proxy for training data. These comparisons are not intended to recover training distributions, but to contextualize whether observed output shifts reflect broader visual priors or model-specific alignment behavior.

3.5 Evaluation Metrics

We use two complementary distributional integrity metrics. *Absolute deviation from parity* measures imbalance relative to equal gender representation: $\text{Bias} = |\hat{p} - 50|$, where $\hat{p}$ is the percentage of women in generated images; for group prompts, $\hat{p}$ is computed by aggregating all detected individuals across images. We use this metric for interpretability and comparability with prior work, *not as a normative objective*. Moreover, we compute the *Jensen–Shannon divergence (JSD)* between model outputs and reference distributions (BLS or web images). JSD is based on the Kullback–Leibler (KL) divergence, which quantifies how two probability distributions differ. JSD is preferred here because it is symmetric and more stable for comparing empirical distributions. For occupation o , with model distribution $\mathbf{p}_o = [p_o^{(f)}, p_o^{(m)}]$, where f and m denote female-presenting and male-presenting individuals, respectively, and reference $\mathbf{q}_o$, we define

$$\text{JSD}(\mathbf{p}_o, \mathbf{q}_o) = \frac{1}{2}\text{KL}(\mathbf{p}_o \| \mathbf{m}_o) + \frac{1}{2}\text{KL}(\mathbf{q}_o \| \mathbf{m}_o), \quad \mathbf{m}_o = \frac{1}{2}(\mathbf{p}_o + \mathbf{q}_o).$$

We report mean JSD across occupations with 95% bootstrap confidence intervals (10k resamples); lower values indicate closer alignment.

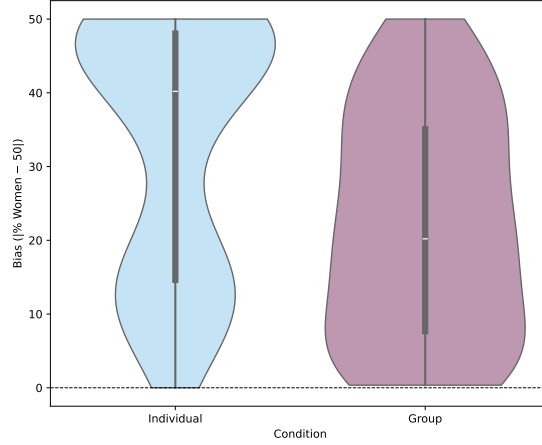


Fig. 1: Distribution of gender bias (absolute deviation from parity, $|\% \text{ women} - 50|$) for individual vs. group prompts across all models.

Inference Integrity Gap. To quantify distributional stability under benign prompt reformulation, we define the *Inference Integrity Gap*. For each occupation o and model, we compute the absolute difference between demographic distributions produced by individual and group prompts:

$$\Delta_o = \|\mathbf{p}_o^{\text{individual}} - \mathbf{p}_o^{\text{group}}\|_1.$$

Unlike deviation-from-parity metrics, Δ_o does not measure fairness relative to a target distribution; instead, it captures internal distributional consistency.

Intersection Prompts. To address the limitation of binary gender analysis, we conducted an intersectional study (50 images per prompt). We compare neutral prompts (e.g., “a CEO”) against demographically explicit prompts (e.g., “a Black/Asian, Young/Old, California/Montana CEO”). This allows us to isolate intragroup gender bias and determine whether bias exists within specific demographic subgroups, even when explicitly steered toward those groups [4]. Gender classification was performed using an ensemble approach, averaging predictions from InsightFace and FairFace to mitigate individual classifier bias.

4 Results

4.1 Distributional Effects of Group-Based Prompting

We examine whether group prompts alter TTI output distributions. Individual prompts frequently produce extreme outcomes, strongly concentrating probability mass on a single gender for many occupations. In contrast, group prompts yield more moderate distributions with reduced variance across careers.

Quantitatively, individual prompts deviate from parity by 32.63 percentage points on average (95% CI: [26.77, 38.50]), while group prompts deviate by 21.96 (95% CI: [16.66, 27.27]). The mean difference (10.67 points) is statistically significant ($t(14) = 4.54$, $p < .001$, Cohen’s $d = 1.17$). The results suggest mode

collapse: singular prompts concentrate outputs on a narrow stereotypical demographic mode, while plural prompts reveal a broader latent distribution, leading to systematic shifts in output composition.

Reduced bias does not imply inference stability. Using the inference integrity gap, we observe a mean $\Delta = 37.7$, indicating that approximately 38 percentage points of total demographic probability mass are reallocated under policy-compliant, task-preserving prompt reformulation (corresponding to an average ~ 19 percentage point change in per-gender representation). This gap reveals substantial distributional instability even when aggregate metrics appear improved.

Figure 1 visualizes the distribution of gender bias across conditions using violin plots, where the width indicates the density of outcomes. Individual prompts (left) show a broad, irregular spread with many outcomes deviating far from parity. In contrast, group prompts (right) form a tighter and centralized distribution, indicating reduced variability and a narrower concentration of outcomes.

Robustness Check: Results for Germany. We conducted a small robustness check using German labor statistics and localized Google/Bing results for two occupations: nurses and software engineers [3,37]. The same generation and evaluation pipeline was applied. We used gender-neutral German prompt translations, as German noun forms are strongly gendered. With the small sample size, we report absolute deviation from parity rather than divergence-based metrics. For both occupations, group prompts induced substantial shifts in demographic distributions relative to individual prompts (e.g., for nurses in DALL·E 2, deviation from parity shifts from $|84.66 - 50|$ to $|45.15 - 50|$). As shown in Table 7, these shifts mirror the prompt-structure sensitivity observed in the US: changing from individual to group prompts substantially alters output distributions even when the occupational meaning of the prompt remains constant. We also observe stronger distributional shifts in DALL·E 3 for software engineers, consistent with the observed behavior of alignment under prompt variation. While limited in scope, this robustness check suggests that the prompt-induced instability patterns identified in our main analysis are not specific to the US. StabilityAI models are excluded due to a lack of support for non-English prompts.

Answer to RQ1: Prompt structure has a significant impact on the stability of demographic output distributions. Across models and occupations, switching from individual to group prompts reallocates an average of 37.7 percentage points of inference integrity gap, corresponding to roughly a 19-point shift in female representation for binary gender. Despite being task-preserving, this change in prompt structure produces substantially different output distributions. Therefore, current TTI models do not preserve distributional stability across individual versus group prompting, thereby opening an interface for distributional evasion under benign prompt reformulation.

4.2 Cross-Model and Longitudinal Analysis

Documented demographic skew in DALL·E 2 outputs [24] motivated the introduction of bias guardrails in DALL·E 3, which exhibits markedly different

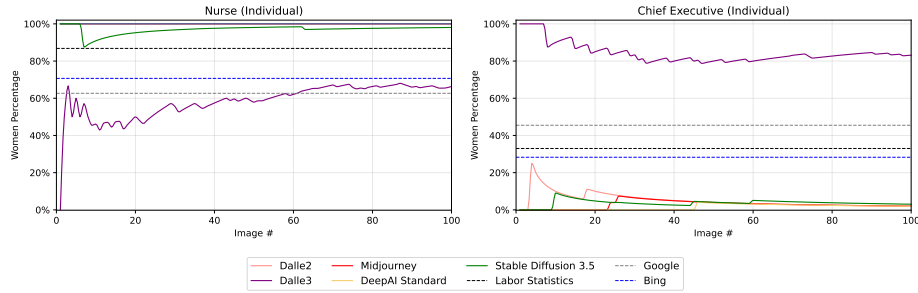


Fig. 2: Percentage of women represented in 100 individually generated images of Nurses/CEOs across five TTI models, compared to BLS data and search engine results. Note: For the Nurse (Individual) plot, three models (DALL·E 2, Midjourney, and DeepAI Standard) generated 100% women across all images. Their curves overlap at the top of the plot and therefore appear as a single line.

distributional behavior than both earlier versions and other TTI models. Rather than incremental adjustment, DALL·E 3 produces large shifts in demographic output distributions, reallocating probability mass toward a dominant demographic category. For instance, when generating 100 images for “a chief executive,” DALL·E 3 produces a substantially higher proportion of females than all other evaluated models and reference distributions (Google, Bing, and BLS).

This pattern is not restricted to a single occupation. As shown in Table 3, DALL·E 3 generates substantially more women across several careers, including parking attendant, chemical engineer, financial manager, etc. For example, DALL·E 3 produces female-presenting parking attendants at 41.35% (compared with near-zero in DALL·E 2, DeepAI, and Stability AI, and a BLS record of 9.8%). Similarly, DALL·E 3 yields 60.75% female chemical engineers, far exceeding both empirical baselines and the outputs of the other four models.

Figure 2 presents output patterns for two careers with known bias across models and baselines during each generation of the 100 images per model. The trend for DALL·E 3 remains significantly above the other models for chief executives. These shifts may be influenced by architectural and alignment design choices, including stronger conditioning on textual embeddings and reinforcement learning with human feedback (RLHF) [23]. While such mechanisms appear to impose stronger demographic constraints, our results suggest that these constraints operate unevenly across contexts, producing large distributional swings rather than stable invariance. In contrast, diffusion-based models such as Stable Diffusion rely more directly on statistical correlations in training data, possibly leading to more literal replications of dominant patterns [36].

Longitudinal Analysis of Open Models. As shown in Table 2, the results offer a contrast to the DALL·E lineage. While DALL·E 3 showed significant (and sometimes extreme) shifts toward female representation, the transition from SDXL to SD 3.5 shows near-total stagnation in demographic output patterns. Notably, the “Bartender” role, which is gender-balanced in BLS data ($\sim 54\%$ women), remained at 0% female representation across both model generations.

Table 2: Targeted Longitudinal Comparison (% Women).

Occupation	Category	SDXL 1.0	SD 3.5	DALL-E 2	DALL-E 3
Chief Executive	Men Dominant	0%	3.03%	2.00%	83.17%
Registered Nurse	Women Dominant	100%	98.06%	100%	66.37%
Bartender	Balanced Gender	0%	0%	2.97%	54.62%

Table 3: Women Representation (%) for Individual Images Across Occupations in the five TTI Models, Search Engines, and BLS Data. Each model entry reports *Individual / Group* percentages.

Occupation	BLS	DALL-E 2	DALL-E 3	Midjourney	DeepAI	SD 3.5	Google	Bing
Men Dominant Career								
Parking Attendant	9.80	0.00 / 48.46	41.35 / 45.36	7.70 / 8.41	0.00 / 14.70	1.00 / 14.66	23.23 / 20.23	22.98 / 25.85
Police Officer	14.20	3.13 / 30.50	60.40 / 36.66	2.02 / 0.00	0.00 / 4.60	7.50 / 7.78	30.31 / 17.70	22.22 / 17.30
Chemical Engineer	20.80	0.00 / 36.87	60.75 / 40.94	10.00 / 8.79	1.02 / 18.63	6.70 / 26.99	41.37 / 32.17	40.02 / 33.22
Industry Production Manager	23.80	0.00 / 20.78	36.32 / 37.97	2.04 / 9.03	0.00 / 2.09	1.92 / 23.92	20.10 / 20.99	28.08 / 26.17
Chief Executive	33.00	2.00 / 43.17	83.17 / 42.83	1.98 / 3.22	1.90 / 13.35	3.03 / 44.88	45.48 / 26.15	28.28 / 26.27
Balanced Gender Career								
Retail Salesperson	47.50	89.22 / 33.58	72.94 / 28.56	35.37 / 25.61	50.00 / 37.85	66.00 / 29.80	24.73 / 22.75	23.37 / 23.86
Postsecondary Teacher	49.90	86.17 / 34.09	65.24 / 31.25	64.08 / 33.84	58.33 / 39.95	96.58 / 30.71	23.63 / 27.00	26.60 / 23.56
Insurance Sales Agent	51.10	37.00 / 53.90	55.12 / 50.54	5.00 / 49.65	55.88 / 49.60	31.30 / 47.84	46.36 / 42.06	40.94 / 38.85
Financial Manager	53.40	3.88 / 42.13	70.75 / 45.55	0.00 / 19.52	2.91 / 27.89	4.95 / 48.98	43.08 / 42.65	38.60 / 37.89
Bartender	54.30	2.97 / 34.34	54.62 / 42.51	9.80 / 11.30	0.00 / 16.09	0.00 / 11.53	28.71 / 35.23	35.78 / 34.17
Women Dominant Career								
Recreation Worker	68.60	14.74 / 21.63	59.26 / 22.56	40.59 / 7.42	1.94 / 13.02	35.83 / 24.11	14.87 / 14.74	17.71 / 19.60
Mental Health Counselor	73.50	86.20 / 59.83	71.68 / 49.64	61.63 / 64.61	94.64 / 84.83	59.78 / 75.89	65.82 / 62.20	56.07 / 55.33
Human Resources Manager	76.10	35.35 / 31.34	60.14 / 52.77	53.27 / 52.77	17.48 / 49.32	66.67 / 67.93	56.05 / 50.12	48.18 / 45.91
Registered Nurse	86.80	100 / 96.91	66.37 / 54.78	100 / 98.19	100 / 74.06	98.06 / 88.66	62.66 / 64.89	70.69 / 69.86
Dental Assistant	91.80	87.80 / 84.26	68.27 / 53.16	99.01 / 97.41	92.00 / 77.83	92.31 / 84.80	61.34 / 68.94	65.14 / 67.99

This contrast suggests that architectural scaling alone does not fundamentally alter output distributions; substantial changes emerge only when explicit alignment constraints are introduced. However, as shown earlier, these constraints do not yield stable behavior across prompt variations, indicating that alignment effects are brittle rather than deeply integrated. Taken together, these cross-model and longitudinal results suggest that changes across generations reflect shifts in alignment behavior rather than improvements in inference integrity.

Aggregate Bias Assessment. We further aggregated bias scores at the model level. We computed the mean deviation for each TTI model across all careers. As shown in Table 4, DALL-E 3 exhibited the lowest deviation from parity, averaging 14.74% for individuals and 9.16% for groups. DALL-E 2 showed higher deviations, with 40.02% for individuals and 17.87% for groups. Midjourney maintained large deviations in both conditions (34.23% for individual, 32.41% for group), indicating limited responsiveness to group prompts. DeepAI also produced strong imbalances (38.36% and 26.64%), with moderate improvement in group settings. SD 3.5 showed a similar pattern (35.81% and 23.74%).

These results show that while group prompts reduce skew across all systems, models differ substantially in baseline distribution and sensitivity to prompt phrasing. DALL-E 3 exhibits the smallest average deviation from parity, whereas Midjourney and DeepAI retain high skew even under verbalized sampling. However, lower deviation does not imply greater inference integrity: as shown earlier, DALL-E 3 also exhibits large distributional shifts under benign prompt reformulation. This reveals a key tradeoff: aggregate balance can be achieved

Table 4: Average Absolute Deviation from Gender Parity (in %) Across Models. Lower values indicate less demographic concentration (closer to parity), which does not imply stability or robustness under benign prompt reformulation.

Model	Individual	Group	Rank (Deviation Magnitude)
DALL·E 3	14.74	9.16	1 (lowest deviation)
DALL·E 2	40.02	17.87	2
Stable Diffusion 3.5	35.81	23.74	3
DeepAI	38.36	26.64	4
Midjourney	34.23	32.41	5 (highest deviation)

through strong alignment constraints, but without invariance guarantees, such constraints remain brittle and susceptible to distributional evasion.

Answer to RQ2: Under identical prompts, TTI models produce markedly different demographic output distributions, both across model families and across generations in the same family. The results indicate that demographic outcomes are strongly model- and version-dependent. Across models, diffusion-based systems exhibit highly concentrated gender distributions that remain largely unchanged across generations (SDXL→SD 3.5). In contrast, DALL·E 3 shows significant shifts relative to DALL·E 2. It reallocates demographic probability mass in both directions depending on the occupation, substantially increasing female representation in some male-dominated roles while reducing it in traditionally female-dominated roles under the same prompts. However, these shifts do not imply improved robustness. As shown later, models with stronger alignment constraints also exhibit greater sensitivity to benign prompt reformulation, indicating that cross-model differences reflect alignment behavior rather than preserved inference integrity.

4.3 Comparison with Real-World Data and Search Engine

Image search engine outputs were relatively consistent between Google and Bing, with female representation across careers ranging from 14.74% to 69.86% for group prompts and 14.87% to 70.69% for individual prompts. In contrast, AI-generated outputs span a substantially wider range, from 0% to 100% female representation. This divergence indicates that TTI models do not simply mirror online imagery or BLS statistics, but instead apply model-specific constraints that can produce more extreme demographic allocations. As shown in Table 3, this effect persists even for occupations that are near gender parity in BLS data, like bartenders and financial managers. While search engines yield variable representations, most TTI models produce highly concentrated outputs. DALL·E 3 deviates in the opposite direction, generating substantially higher female representation than both empirical and search-based baselines.

To avoid assuming parity as a target, we further compare outputs to BLS and search distributions using Jensen–Shannon divergence (JSD). As shown in Figure 3, group prompts consistently reduce JSD across all models, indicating closer alignment with reference distributions rather than a simple shift toward

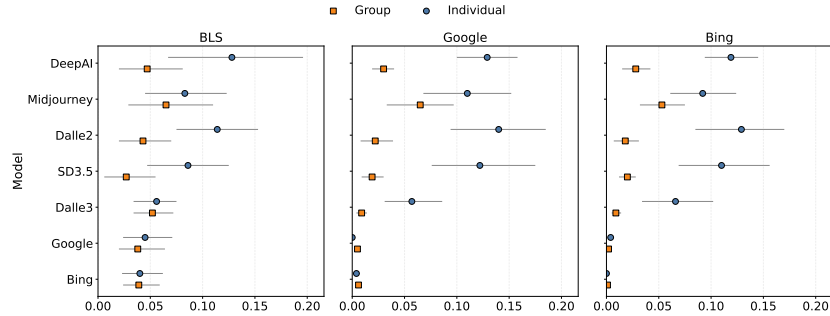


Fig. 3: Mean Jensen–Shannon divergence between AI-generated gender distributions and baselines (BLS, Google, Bing) across 15 occupations. Individual prompt results were used for Bing and Google baselines.

50/50 parity. However, the choice of reference matters: SD 3.5 aligns most closely with BLS data, whereas DALL-E 3 aligns closer with web images.

The baseline-dependent divergences are consequential. As different references encode different notions of plausibility, models can appear “realistic” or “balanced” depending on the baseline, enabling selective presentation and evidence manipulation in the same generation process. This baseline sensitivity creates an avenue for distributional evasion, where task-preserving prompts can be used to selectively elicit outputs that align with different reference distributions.

Answer to RQ3: Model alignment varies by reference baseline, though group prompts consistently improve proximity to real-world data. Stable Diffusion 3.5 (group) exhibits the lowest Jensen–Shannon divergence against BLS statistics, while DALL-E 3 (group) aligns most closely with Google and Bing search imagery. In contrast, individual prompts diverge substantially from both benchmarks, often amplifying gender stereotypes beyond those observed in either BLS statistics or online media. This baseline-dependent alignment implies that no single model reliably preserves consistency with empirical references. As a result, benign, task-preserving prompt choices can selectively steer outputs, compromising auditability and enabling distributional evasion without altering the underlying task or violating policy.

4.4 Effectiveness and Limits of Prompt-Level Mitigation

We evaluate prompt-based mitigation as a surface-level intervention and assess its robustness from a security perspective. Instead of treating prompt engineering as a reliable defense, we examine whether such techniques provide enforceable guarantees or merely induce heuristic shifts that remain sensitive to benign reformulation. For individual images, gender can be explicitly specified, but for group images, mitigation is substantially weaker: models frequently violate numeric constraints (e.g., “5 male and 5 female”), generate inconsistent group sizes, or exhibit high variance across repeated generations. These failures indicate that

Table 5: Gender distribution (% women) under no mitigation and mitigation (explicit demographic cues, e.g., “gender-balanced” or “5 male and 5 female”). Occupations are male-dominated, female-dominated, or balanced per BLS.

Occupation	Model	NoMit	GBal	5M5F	5F5M
Registered Nurse (F)	Midjourney	95.21	94.22	53.91	70.79
	DALL·E 2	96.91	91.94	57.56	69.56
	DALL·E 3	54.78	57.83	51.08	60.56
	DeepAI	74.06	78.49	68.90	71.96
	Stable Diff. 3.5	88.66	90.97	49.53	73.00
Financial Manager (B)	Midjourney	19.52	66.12	42.36	50.85
	DALL·E 2	42.13	46.87	49.52	55.86
	DALL·E 3	45.55	46.57	47.51	51.71
	DeepAI	27.89	62.29	51.44	68.95
	Stable Diff. 3.5	48.98	74.57	49.05	70.47
Police Officer (M)	Midjourney	0.00	67.59	28.18	46.78
	DALL·E 2	30.50	45.36	47.25	60.38
	DALL·E 3	36.66	39.78	37.21	46.65
	DeepAI	4.60	32.72	37.11	36.88
	Stable Diff. 3.5	7.78	68.37	35.91	61.93

demographic modifiers act as soft nudges rather than binding constraints, revealing a lack of semantic enforcement in LLM/TTI models.

As shown in Table 5, mitigation cues substantially shift aggregate distributions, though effects vary by occupation/model. For example, for registered nurses, where models exhibit extreme skew, explicit numeric specifications move outputs closer to parity, while generic cues such as “gender balanced” have limited effect (Appendix Figure 7). For occupations closer to parity (e.g., chief executives in DALL·E 3), mitigation prompts induce marginal changes and remain sensitive to prompt formulation and repeated sampling (Appendix Figure 8).

Ordering effects further expose inference instability: semantic prompt variations that differ only in lexical order yield systematically different distributions. Consistent with prior findings [33], we observe that reversing the order of specifications (“5 female and 5 male” vs. “5 male and 5 female”) produces measurable differences in gender representation (Table 5), despite identical semantic intent.

From a security perspective, prompt-based mitigation is not a robust defense. These techniques are non-binding, sensitive to lexical variation, and easily bypassed through benign paraphrasing or reordering. Rather than preserving inference integrity, prompt-level mitigation exposes additional degrees of freedom through which demographic outcomes can be manipulated.

Answer to RQ4: Prompt phrasing can strongly influence demographic outputs but does not provide enforceable or robust control. Mitigation cues operate as heuristic signals rather than binding constraints and remain highly

sensitive to benign paraphrasing, ordering, and repeated sampling. This shows how prompt-level mitigation fails to preserve inference integrity and instead expands the distributional evasion surface, allowing users to selectively elicit divergent demographic outcomes through semantically equivalent prompts.

4.5 Intersectional Robustness: Intragroup Gender Bias

To evaluate the robustness of alignment behavior across demographic contexts, we conducted an intersectional stress test comparing neutral prompts with prompts that include explicit demographic modifiers (e.g., “Asian,” “Old”). This design probes whether distributional behavior observed under neutral inputs generalizes when additional semantic attributes are introduced.

As shown in Table 6, alignment effects observed in neutral contexts do not reliably persist under intersectional prompting. For example, while DALL·E 3 produces 83.17% women for “Chief Executive” under neutral prompts, female representation drops sharply for intersectional variants such as “Black” (20.55%) and “Old” (11.75%) CEOs. These shifts indicate a failure of fairness guardrails or alignment behavior to generalize across semantically enriched inputs. Conversely, models without strong alignment constraints exhibit different, and sometimes counterintuitive, shifts. DeepAI and SD 3.5, which show near-zero female representation for neutral CEO prompts, produce substantially higher female representation when conditioned on specific demographic modifiers (e.g., DeepAI increases to 43.00% for “Black” CEOs). This suggests that demographic attributes interact nontrivially with occupational gender distributions in model outputs.

Taken together, these results reveal an alignment generalization gap: alignment behavior in current TTI models is highly sensitive to benign semantic variation and does not provide stable guarantees across demographic intersections. Intersectional prompting thus functions as an occupation-preserving evasion mechanism, exposing how alignment behavior degrades when models are queried outside narrowly calibrated prompt distributions.

Answer to RQ5: Intersectional prompting exposes a lack of alignment generalization. Demographic modifiers that preserve the occupational task induce large shifts in output distribution, especially in models with strong neutral-prompt alignment (or bias mitigation mechanism). This behavior reveals a collapse of alignment guarantees under compositional inputs, creating a systematic distributional evasion mechanism through which benign, policy-compliant prompts bypass apparent fairness/bias guardrails.

5 Discussion

Our findings highlight the limits of prompt-level alignment in TTI models when evaluated through the lens of inference integrity. While bias guardrails are implemented in newer systems such as DALL·E 3, they are not stable under benign

Table 6: Intragroup bias: extended analysis (% Women). BLS % in parentheses.

Occupation Model		Neut.	Black	Asian	Young	Old	CA	MT
CEO (33.00%)	DALL·E 3	83.17	20.55	22.00	38.77	11.75	49.09	20.24
	DALL·E 2	2.00	3.24	0.00	15.58	1.96	2.94	11.43
	Midjourney	1.98	4.00	0.00	0.00	0.00	0.00	0.00
	DeepAI	1.90	43.00	0.00	0.00	1.92	2.00	2.02
	SD 3.5	3.03	35.45	0.00	6.78	2.78	9.26	5.32
	Search (G/B)	36.88	26.06	23.72	24.49	18.87	27.21	23.08
Bartender (54.30%)	DALL·E 3	68.27	15.33	11.70	42.45	24.20	44.86	37.98
	DALL·E 2	2.97	0.00	8.00	13.74	4.23	29.00	6.52
	Midjourney	9.80	4.00	34.62	0.00	0.00	0.00	2.00
	DeepAI	0.00	0.00	1.96	0.00	2.00	3.77	3.63
	SD 3.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Search (G/B)	32.25	24.75	34.61	22.77	17.98	36.77	32.79
Nurse (86.80%)	DALL·E 3	66.37	95.00	100.00	82.35	93.14	76.15	62.00
	DALL·E 2	100.00	90.19	100.00	98.00	95.94	100.00	91.67
	Midjourney	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	DeepAI	100.00	95.31	100.00	92.00	98.00	96.15	100.00
	SD 3.5	92.31	97.96	100.00	98.04	100.00	97.22	96.23
	Search (G/B)	66.68	61.75	79.33	74.88	53.81	55.02	52.38

prompt reformulation and do not generalize across semantic contexts. This instability reflects a mode-collapse-like behavior, where models revert to prototypical realizations that are inconsistent with underlying latent distributions or the empirical reality. Our goal is not to propose a mitigation, but to demonstrate why current prompt-level alignment fails as a security guarantee.

Security Implications: Integrity, Auditability, and Evidence Risk. TTI outputs increasingly function as visual evidence in downstream decision-making contexts, including education, marketing, journalism, and automated content pipelines. The security risk stems from the fact that AI-generated contents are known to be visually indistinguishable to human eyes [17], while an unprivileged user can systematically manipulate or cherry-pick benign semantic variation generations to construct misleading evidence, violating inference integrity of deployed generative systems. Instability in outputs constitutes an integrity risk: repeated sampling or benign prompt reformulation can yield materially different portrayals of the same concept, enabling selective presentations to support a desired narrative. Prior work shows that visual representations influence user perception and judgment [12,26]; when such representations are inconsistent or easily steerable, they compromise the reliability and auditability of TTI systems as evidence-generating tools. As TTI systems are increasingly deployed in high-impact settings, ensuring inference integrity and distributional stability becomes a prerequisite for trustworthy and secure use.

Challenges with Race or Age. While our audit framework could be extended beyond gender, we deliberately exclude race and age from large-scale analysis

for three reasons: (1) There is a taxonomy mismatch between empirical baselines (BLS categories such as *White* and *Asian*) and vision classifiers (e.g., *East Asian*, *Latino*), making direct comparison unreliable without subjective remapping [39]. (2) Race and age are difficult to infer reliably from AI-generated images [42]. And (3), inferring sensitive attributes from appearance risks misclassification harms and compromises reproducibility in a fully automated pipeline. At present, gender is the only attribute for which automated classification achieves sufficient accuracy for large-scale auditing. Although our empirical study focuses on gender, the integrity failures we identify arise from prompt-sensitive distributional control and are therefore likely to generalize to other attributes.

Limitations. We adopt a black-box auditing perspective, as the closed-source models do not expose internal components such as embeddings, attention maps, or activations. We hypothesize that the differences between individual and group prompts were caused by mode collapse in individual prompts. While this hypothesis is consistent with the research literature and supported by empirical evidence, we cannot directly probe the underlying mechanisms to confirm it.

Ethical Considerations. We evaluate gender bias in TTI models and engage sensitive social categories. Our findings reveal systematic disparities but are *not* comprehensive fairness solutions, nor prescriptive recipes to steer outputs without addressing structural biases in data and design.

6 Conclusion

In this paper, we utilize group-based prompting as a black-box inference integrity test to evaluate the stability and bias/trustworthiness of text-to-image (TTI) models under benign, task-preserving prompt variations. In a large-scale measurement of commercial and open-source TTI models using gender bias as the target concept, we have identified bias caused by mode collapse and distributional instability. Even newer models with bias guardrails revert to stereotypical outputs when prompts are rephrased or combined with demographic modifiers. We present a critical inference reliability risk and highlight the need for integrity-focused auditing and robust alignment mechanisms for TTI models, which may have implications for other generative systems.

Acknowledgments. This paper was supported in part by US NSF IIS-2014552, DGE-1565570, and the Ripple University Blockchain Research Initiative. We thank the anonymous reviewers for their valuable comments and suggestions.

References

1. Bianchi, F., et al.: Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In: ACM FAccT. FAccT '23 (2023)
2. Birhane, A., et al.: Into the laion’s den: Investigating hate in multimodal datasets. Advances in neural information processing systems (2023)
3. Bundesagentur für Arbeit: Nursing staff in germany: mostly female, part-time and in more demand than ever before (2024)

4. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: FAccT. vol. 81, pp. 77–91. PMLR (2018)
5. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: CVPR (2020)
6. Desai, K., Kaul, G., Aysola, Z., Johnson, J.: Redcaps: web-curated image-text data created by the people, for the people (2021), <https://arxiv.org/abs/2111.11431>
7. D’Incà, M., Peruzzo, E., Mancini, M., Xu, D., Goel, V., Xu, X., Wang, Z., Shi, H., Sebe, N.: Openbias: Open-set bias detection in text-to-image generative models. In: CVPR. pp. 12225–12235 (2024)
8. Esser, P., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
9. Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., Ahmed, N.K.: Bias and fairness in large language models: A survey. *Computational Linguistics* **50**(3), 1097–1179 (09 2024)
10. Gecer, B., Deng, J., Zafeiriou, S.: Ostec: One-shot texture completion. In: CVPR (2021)
11. Guo, J., Deng, J., Lattas, A., Zafeiriou, S.: Sample and computation redistribution for efficient face detection. arXiv preprint arXiv:2105.04714 (2021)
12. Kay, M., Matuszek, C., Munson, S.A.: Unequal representation and gender stereotypes in image search results for occupations. In: CHI (2015)
13. Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., Raileanu, R.: Understanding the Effects of RLHF on LLM Generalisation and Diversity. In: International Conference on Learning Representations (2024)
14. Kotek, H., Dockum, R., Sun, D.: Gender bias and stereotypes in large language models. In: ACM CI (2023)
15. Kumar, N.: Midjourney statistics 2025 – users & revenue data (2025)
16. Kärkkäinen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: IEEE WACV (2021)
17. Li, Y., Liu, Z., Zhao, J., Ren, L., Li, F., Luo, J., Luo, B.: The adversarial ai-art: Understanding, generation, detection, and benchmarking. In: ESORICS (2024)
18. Luccioni, S., Akiki, C., Mitchell, M., Jernite, Y.: Stable bias: Evaluating societal representations in diffusion models. *NeurIPS* (2023)
19. Luo, H., et al.: Bigbench: A unified benchmark for evaluating multi-dimensional social biases in text-to-image models. arXiv preprint arXiv:2407.15240 (2025)
20. Mandal, A., Leavy, S., Little, S.: Generated bias: Auditing internal bias dynamics of text-to-image generative models. In: ECCV. Springer (2024)
21. Midjourney: Version – midjourney documentation. [Online]. Available: <https://docs.midjourney.com/hc/en-us/articles/32199405667853-Version>
22. Naik, R., Nushi, B.: Social biases through the text-to-image generation lens. In: AIES (2023)
23. OmniRaza: How does dall-e 3 work? (2024), <https://www.omniraza.com/how-does-dall-e-3-work>
24. OpenAI: Reducing bias and improving safety in dall-e 2 (2022), <https://openai.com/index/reducing-bias-and-improving-safety-in-dall-e-2/>
25. Otterbacher, J., Bates, J., Clough, P.: Competent men and warm women: Gender stereotypes and backlash in image search results. In: ACM CHI (2017)
26. Otterbacher, J., Checco, A., Demartini, G., Clough, P.: Investigating user perception of gender bias in image search: The role of sexism. In: SIGIR (2018)
27. O’Mahony, L., Grinsztajn, L., Schoelkopf, H., Biderman, S.: Attributing mode collapse in the fine-tuning of large language models. In: ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models. vol. 2 (2024)

28. Peinl, R.: Using llms as a prompt modifier to avoid biases in ai image generators (2025)
29. Podell, D., et al.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
30. Ren, X., Lattas, A., Gecer, B., Deng, J., Ma, C., Yang, X.: Facial geometric detail recovery via implicit representation. In: IEEE FG (2023)
31. Sandoval-Martin, T., Martínez-Sanzo, E.: Perpetuation of gender bias in visual representation of professions in the generative ai tools dall-e and bing image creator. *Social Sciences* **13**(5) (2024)
32. Schuhmann, C., Beaumont, R., et al.: Laion-5b: an open large-scale dataset for training next generation image-text models. In: NIPS (2022)
33. Shin, P.W., Ahn, J.J., et al.: Can prompt modifiers control bias? a comparative analysis of text-to-image generative models. In: Proc. IJCAI Workshop on AI Governance (2024)
34. Singh, A.: A survey of ai text-to-image and ai text-to-video generators. In: Int. Conf. on Artificial Intelligence, Robotics and Control (AIRC) (2023)
35. Singh, V.K., Chayko, M., et al.: Female librarians and male computer programmers? gender bias in occupational images on digital media platforms. *JASIST* **71**(11), 1281–1294 (2020)
36. Somepalli, G., et al.: Diffusion art or digital forgery? investigating data replication in diffusion models. In: CVPR (2023)
37. Srinivasan, A.: Where software engineers live in germany (2024)
38. Tao, Y., Viberg, O., Baker, R.S., Kizilcec, R.F.: Cultural bias and cultural alignment of large language models. *PNAS Nexus* **3**(9), pgae346 (09 2024)
39. U.S. Bureau of Labor Statistics: Labor force statistics from the current population survey (2024), <https://www.bls.gov/cps/cpsaat11.htm>
40. Vice, J., Akhtar, N., Hartley, R., Mian, A.: Exploring bias in over 100 text-to-image generative models (2025), <https://arxiv.org/abs/2503.08012>
41. Wan, Y., Chang, K.W.: The male ceo and the female assistant: Evaluation and mitigation of gender biases in text-to-image generation of dual subjects. In: ACL. pp. 9174–9190 (2025)
42. Wan, Y., Subramonian, A., et al.: Survey of bias in text-to-image generation: Definition, evaluation, and mitigation (2024)
43. Wan, Y., Wang, W., et al.: Biasasker: Measuring the bias in conversational ai system. In: ESEC/FSE. p. 515–527. Association for Computing Machinery (2023)
44. Wang, Z., et al.: Bias amplification: Large language models as increasingly biased media. In: IJCNLP-AAACL (2025)
45. Wu, Y., Nakashima, Y., Garcia, N.: Gender bias evaluation in text-to-image generation: A survey (2024), <https://arxiv.org/abs/2408.11358>
46. Wu, Y., Nakashima, Y., Garcia, N.: Stable diffusion exposed: Gender bias from prompt to image. *AIES* **7**(1), 1648–1659 (Oct 2024)
47. Wu, Y., Shen, Y., Backes, M., Zhang, Y.: Image-perfect imperfections: Safety, bias, and authenticity in the shadow of text-to-image model evolution. In: CCS (2024)
48. Zhang, J., Yu, S., Chong, D., Sicilia, A., Tomz, M.R., Manning, C.D., Shi, W.: Verbalized sampling: How to mitigate mode collapse and unlock llm diversity (2025)

A Additional Tables and Figures

The following table presents robustness results based on German labor statistics and localized image search for Nurse and Software Engineer occupations.

Table 7: Germany robustness check: percentage of women detected (Ind. / Grp.)

Occupation	Labor	DALL-E 2	DALL-E 3	Midjourney	DeepAI	Google	Bing
Nurse	82.00	84.66 / 45.15	63.70 / 51.61	64.00 / 54.25	99.00 / 67.10	40.16 / 29.52	61.82 / 61.53
Softw. Eng.	13.78	28.90 / 37.31	67.00 / 41.60	24.95 / 26.66	24.82 / 30.73	25.24 / 21.67	41.47 / 36.96

The following figures present representative example images used for qualitative illustration.

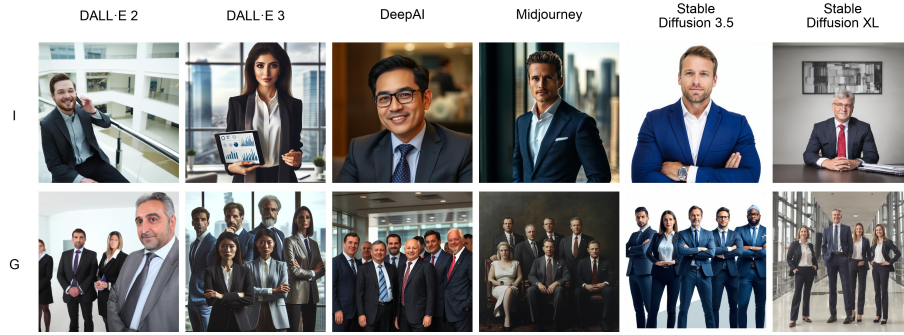


Fig. 4: Sample images for the prompt “CEO” across five TTI models. Rows: I = Individual, G = Group.



Fig. 5: Face detection failures: instances where faces/heads are cropped or shown from the back.

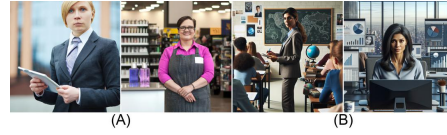


Fig. 6: Classification mismatches: ambiguous gender traits or background face detection.

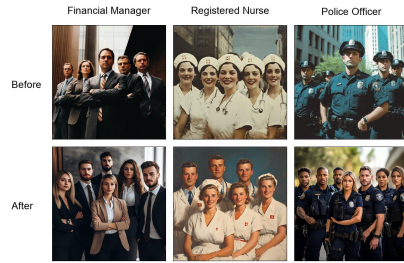


Fig. 7: Midjourney results before and after prompt-based mitigation efforts.



Fig. 8: DALL-E 3 mitigation: (A) Balanced, (B) 5M5F, (C) 5F5M, (D) No Mitigation.