Operationalizing Fairness in Text-to-Image Models: A Survey of Bias, Fairness Audits and Mitigation Strategies

Megan Smith[1*] **Venkatesh Thirugnana Sambandham**[1*] **Florian Richter**[2]
Laura Crompton[1] **Matthias Uhl**[3] **Torsten Schön**[1]

[1]**AImotion Bavaria, Technische Hochschule Ingolstadt, Ingolstadt, Germany**
[2]**School of Transformation and Sustainability, Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany**
[3]**Chair of Economic and Social Ethics, University of Hohenheim, Stuttgart, Germany**

{megan.smith, venkatesh.thirugnanasambandham}@thi.de
{laura.crompton, torsten.schoen}@thi.de
florian.richter@ku.de matthias.uhl@uni-hohenheim.de

Abstract

Text-to-Image (T2I) generation models have been widely adopted across various industries, yet are criticized for frequently exhibiting societal stereotypes. While a growing body of research has emerged to evaluate and mitigate these biases, the field at present contends with conceptual ambiguity, for example terms like "bias" and "fairness" are not always clearly distinguished and often lack clear operational definitions. This paper provides a comprehensive systematic review of T2I fairness literature, organizing existing work into a taxonomy of bias types and fairness notions. We critically assess the gap between "target fairness" (normative ideals in T2I outputs) and "threshold fairness" (normative standards with actionable decision rules). Furthermore, we survey the landscape of mitigation strategies, ranging from prompt engineering to diffusion process manipulation. We conclude by proposing a new framework for operationalizing fairness that moves beyond descriptive metrics towards rigorous, target-based testing, offering an approach for more accountable generative AI development.

Keywords: Text-to-Image, Fairness, Diffusion, Flow Matching, Bias

1. Introduction

Text-to-Image (T2I) foundation models have reshaped digital content creation. They demonstrate capabilities in synthesizing high-fidelity visual content from natural language prompts (Ho et al., 2020; Rombach et al., 2022). However, as these models are integrated into increasingly automated workflows, their tendency to reproduce and amplify societal stereotypes presents a critical challenge to their safe deployment (Bianchi et al., 2023; Naik and Nushi, 2023). For example, neutral prompts often yield images heavily skewed towards specific demographics, while cultural depictions frequently suffer from reductive caricatures (Cheong et al., 2024). This poses a potential for harm if image outputs shift user perceptions of what is typical or representative, which user studies show can alter beliefs about real-world group distributions (Guilbeault et al., 2024; Kay et al., 2015).

In response, the research community has produced a wealth of methodologies to evaluate (Cheong et al., 2024; Mack et al., 2024) and mitigate (Kim et al., 2025; Gandikota et al.,

∗. Equal contributions.

© 2026 .

2024b) these issues. Yet, the field shows signs of nominative fragmentation. For example, researchers sometimes conflate "bias" (the statistical observation of disparity) with "fairness" (the normative goal of equitable representation). This ambiguity presents a potential barrier to alignment: while current methodologies have become adept at detecting representational failures, they lack a consistent framework for defining and enforcing fairness. As a result, many current evaluation and mitigation strategies function as temporary patches rather than structural solutions.

While prior surveys on bias in T2I models (Wan et al., 2024; Elsharif et al., 2025) have extensively cataloged existing definitions and metrics, our work departs from this by establishing a normative framework that aims to bridge the gap between descriptive audits and actionable enforcement. In this work, we identify a gap between **Target Fairness** (setting a normative ideal for what is considered fair) and **Threshold Fairness** (the definition of precise, actionable thresholds for model behavior). In addition, we identify contexts where T2I research would benefit more from aiming toward threshold fairness, and we argue that without moving towards the latter, T2I models risk not fully aligning with the normative standards expected of foundation models.

1.1. Contributions

This review aims to organize the inconsistencies in the T2I fairness landscape into a framework for evaluation and alignment:

- **The Target-to-Threshold Framework:** We introduce a theoretical distinction between *Target Fairness* (normative ideal) and *Threshold Fairness* (normative enforcement). We demonstrate that the majority of current literature is focused on Target Fairness or bias, resulting in audits that quantify bias or measure towards a norm without providing the defined stopping conditions necessary for robust alignment.
- **A Unified Taxonomy of Representational Harms:** We systematize terminology across computer vision and social computing, distinguishing between six distinct forms of bias (including amplification, composition, and censorship bias). We delineate operational fairness notions: demographic parity, proportional representation, and performance parity.
- **Critical Analysis of Mitigation Strategies:** We conduct an extensive review of state-of-the-art interventions, ranging from inference-time prompt engineering to model fine-tuning and latent manipulation. We analyze these methods on their ability to shift distributions and identify trade-offs that currently limit their effectiveness.

The remainder of this paper is organized as follows. Section 2 briefly reviews the technical foundations of diffusion based T2I generation pipeline. In Section 3, we systematize the existing literature into a unified taxonomy of bias types and fairness notions, and introduce our proposed framework for fairness operationalization. Section 4 surveys the landscape of current auditing and evaluation methodologies. Finally, Section 5 categorizes mitigation strategies based on their specific intervention points within the generative pipeline.

2. Foundations: The Roots of Bias in the T2I Pipeline

State-of-the-art T2I frameworks, encompassing both Diffusion (Ho et al., 2020; Rombach et al., 2022) and Flow-Matching (Esser et al. (2024)) paradigms, converge on a shared three-stage architecture. While the generative mechanics differ (probabilistic denoising vs. deterministic velocity fields), the sources of bias are structurally identical across both kinds.

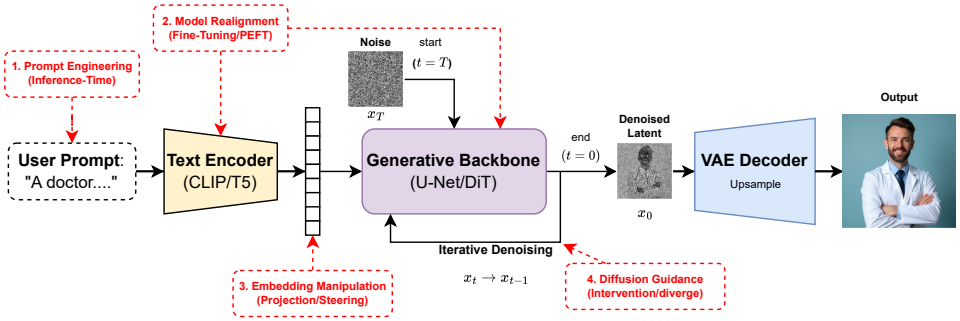


Figure 1: Overview of the inference pipeline in T2I generation. We map four mitigation strategies discussed in Section 5 to their respective intervention points: (1) Prompt modifications at the input, (2) Weight updates via fine-tuning, (3) Geometric Projection in the embedding space, and (4) Dynamic guidance during the denoising loop.

We analyze this pipeline through three critical components (visualized in Fig. 1): the text encoder, the generative backbone, and the latent compression model.

Text Conditioning The input prompt is processed by a pre-trained text encoder, such as CLIP (Radford et al., 2021) or T5 (Raffel et al., 2020). These encoders project natural language into high-dimensional embeddings that act as the primary control signal. Because these encoders are often frozen and pre-trained on internet-scale data, they serve as the first entry point for representational bias. If the encoder associates "doctor" primarily with male embeddings, the downstream generation is restricted before it even begins, limiting the model's ability to interpret diverse prompts correctly.

The Generative Backbone The core network, whether a UNet (Ronneberger et al., 2015) or a Diffusion Transformer (DiT) (Peebles and Xie, 2023), learns to approximate the probability distribution of the training dataset. This component is responsible for bias amplification (Hakemi et al., 2025; Chen et al., 2024): it does not merely reproduce the statistical skews of the training data (e.g., stereotypes regarding race or gender) but often exaggerates them to minimize training objective. This results in distribution sharpening, where the model defaults to the most probable demographic representation for a given class.

Latent Space Operations To maximize efficiency, models like Stable Diffusion operate in a compressed latent space via a perceptual autoencoder (VAE) (Rombach et al., 2022). We identify this abstraction layer as a bias bottleneck. The VAE creates a lossy compression optimized for features dominant in the training data. Consequently, fine-grained details of underrepresented groups, such as specific skin tones, non-Western facial structures, or cultural artifacts may be discarded during compression. Once lost in the latent space, these features cannot be recovered by the decoder, enforcing a homogenized "default" appearance regardless of the prompt.

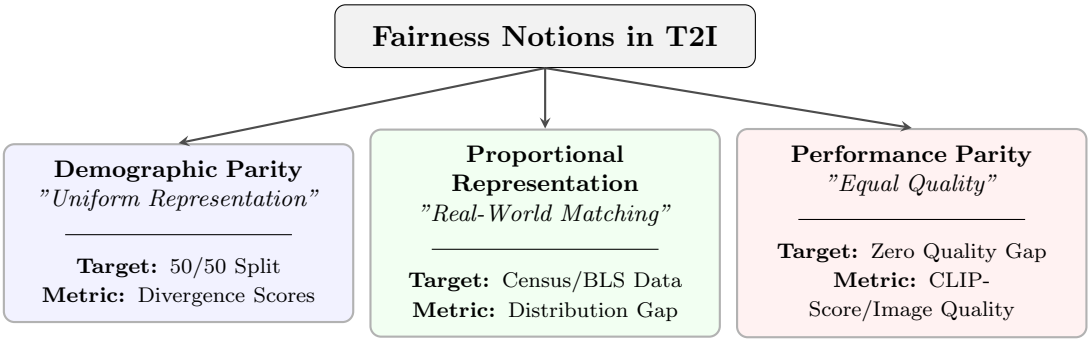


Figure 2: A Unified Taxonomy of Fairness Notions. Objectives are classified into three families: enforcing uniformity (Demographic Parity), mirroring reality (Proportional Representation), or ensuring capability invariance (Performance Parity).

3. Defining Framework of Bias and Fairness Types

This section outlines the terminology, definitions, and assumptions surrounding fairness and bias across our survey of T2I research. Our aim is to normalize the terms used across the literature to support the results and discussion that follow.

3.1. A Unified Taxonomy of Fairness Notions

The T2I literature lacks a singular definition of fairness. Instead, we identify three distinct operational families (visualized in Figure 2), each defined by a specific normative target.

Demographic Parity (Enforcing Uniformity) This notion operationalizes fairness as equal representation, aiming for a uniform distribution (e.g., 50/50 gender split) regardless of real-world statistics. Works such as Hall et al. (2023), Zhang et al. (2023), and Kim et al. (2025) explicitly set a uniform target and measure deviations. Zhang et al. (2024) evaluates demographic parity across groups relative to an explicit uniform target distribution.

Proportional Representation (Mirroring Reality) In contrast to uniformity, this notion aligns model outputs with external real-world baselines, such as U.S. Bureau of Labor Statistics (BLS) or census data. Studies consistently find that T2I models deviate significantly from these baselines, often under-representing women and minorities in high-status occupations (Bianchi et al., 2023; Cheong et al., 2024; Sun et al., 2024). For instance, Naik and Nushi (2023) found that only 8 of 43 occupations generated by T2I models fell within a $\pm 5\%$ band of actual US workforce statistics.

Performance Parity (Quality Invariance) This notion shifts focus from representation counts to task capability, demanding that the model perform equally well (e.g., in image fidelity or prompt alignment) across demographic groups. Metrics include CLIPScore (Hessel et al., 2021) disparities (Bakr et al., 2023; Lee et al., 2023), VQA accuracy gaps (Luccioni et al., 2023), and counterfactual resolution rates (Hall et al., 2023). A critical gap in this domain is the lack of tolerance bands, as most studies implicitly assume a "zero gap" target without propagating uncertainty, making it difficult to distinguish statistical noise from systematic unfairness.

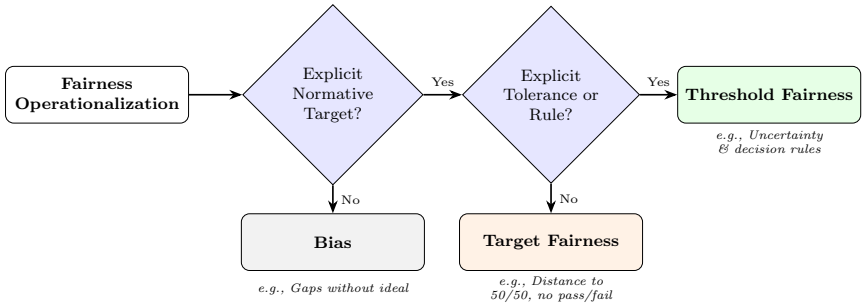


Figure 3: **The Bias-to-Threshold Fairness Framework.** We classify fairness operationalizations based on two criteria: the presence of a normative target (e.g., population parity) and the presence of an explicit decision rule (e.g., tolerance bands).

3.2. A Framework for Bias and Fairness Operationalization

Our survey uncovered a tension between fairness declarations and their operationalization. Drawing from Ferrara et al. (2025) and Corbett-Davies et al. (2023), which argue that rigorous fairness claims require both a clearly specified normative target and an explicit decision rule (rather than descriptive gap reporting alone), we propose a taxonomy to normalize our audit of the T2I fairness landscape. We visualize this classification framework in Figure 3, which categorizes methodologies based on their level of operational rigor.

Target Fairness declares a normative target or ideal (such as demographic parity or proportional representation) and measures the distance to that target. However, it fails to specify an acceptance threshold or decision rule. Consequently, it benchmarks against an ideal without determining actual compliance or establishing a pass/fail condition. Across our survey, this was one of the more common fairness type, with examples including (Sandoval-Martin and Martínez-Sanzo, 2024; Cho et al., 2023; Naik and Nushi, 2023; Luccioni et al., 2023; Zhang et al., 2023; Teo et al., 2024; Hall et al., 2023; Cheong et al., 2024).

Threshold Fairness specifies a normative target alongside an acceptance region, tolerance band, or maximum allowable deviation from that target. Ideally, it employs uncertainty-aware inference to determine if observed disparities are sufficiently small to be deemed "fair", and thus for some evaluation spaces is a more appropriate fairness operationalization. The clearest examples that meet this criteria include Jin et al. (2024), and Cheng et al. (2025). Works that satisfy nearly all requirements include (Jung et al., 2025; Shen et al., 2023; Friedrich et al., 2025), all of which define or optimize toward a fairness target but stop short of specifying a concrete acceptance band, fairness cutoff, or pass/fail compliance rule.

3.3. Bias types

In this survey, bias refers to descriptive measurements of outcomes in T2I outputs which favors some groups or attributes over others, and is represented as skews, distribution shifts, and quality gaps. We group the biases identified in our survey into the categories below.

Representation Bias The output distributions favor (or disfavor) some demographic groups over others. For example, in Sandoval-Martin and Martínez-Sanzo (2024) a model

which generates more men/women despite a neutral occupation prompt. Other examples include (Bakr et al., 2023; Wang et al., 2023; D'Inca et al., 2024; Zhang et al., 2023; Luccioni et al., 2023; Cho et al., 2023).

Stereotyping Bias Groups are linked to roles, traits, settings, visual cues (such as clothing, countenance, posture, economic status) when not prompted. For example, in Bianchi et al. (2023), prompting a system with "a person stealing" generates dark skinned features, unfairly linking criminality to a specific racialized appearance despite the racially neutral prompt. Other examples include (Zhang et al., 2023; Cho et al., 2023; D'Inca et al., 2024; Wang et al., 2023; Lee et al., 2023; Sun et al., 2024; Luccioni et al., 2023).

Amplification Bias The model shifts demographic or attribute proportions away from those observed in the training data. In Seshadri et al. (2024), prompting the system with "engineer" yields about 10% women, despite 25% of women making up the training set. Other examples include (Friedrich et al., 2025; Bianchi et al., 2023; Wu et al., 2025; Gengler, 2024).

Composition Bias In multi-attribute scenes (with subject, object, action, etc.) certain subjects systematically map to biased or unwanted roles/positions. This is seen in Malakouti and Kovashka, in which prompts like "boy feeding woman" revert or "collapse" to the more expected role (woman feeding boy.) Other examples include Bakr et al. (2023).

Censorship Bias Safety/filters selectively suppress or alter content for certain identities or cultural markers. This is seen in Mack et al. (2024) when a user prompted "a person with bipolar disorder, photo," and the system returned: "...this request may not follow our content policy." Other examples include (Ghosh et al., 2025; Baxter, 2023).

Fidelity Bias Task correctness/accuracy differs by group (mis-recognition/misalignment despite equivalent prompts). This can be seen in Baxter (2023) where errors like "White torsos with Black limbs" are generated, diverging from the original prompt.

In our survey, we define bias reports disparities as a measurement (for example, representation gaps) without specifying a normative ideal or pass/fail rule. From this perspective, we treat works which declare a fairness investigation (but do not employ a normative or threshold rule) as investigating bias, not fairness (Lee et al., 2023; D'Inca et al., 2024).

4. Evaluation Strategies

T2I bias and fairness evaluation is dominated by three evaluation spaces that trade scale against interpretive nuance. **Manual audits** employ user studies and author-led counts and ground findings in real-world statistics (proportional representation) and surface bias types that automated tools often miss, such as composition and censorship bias, but do not scale well (Mack et al., 2024; Qadri et al., 2023; Szymański et al., 2025; Sun et al., 2024; Sandoval-Martin and Martínez-Sanzo, 2024). **Automated audits** utilize attribute classifiers, embedding analyses, and VQA/caption tools and extend coverage of gender, race, and occupation at scale, typically operationalizing fairness as demographic parity without a well-justified target, though threshold fairness does appear in a minority of works (AlDahoul et al., 2025; Cheong et al., 2024; D'Inca et al., 2024; Zhang et al., 2023, 2024; Naik and Nushi, 2023; Friedrich et al., 2025; Sreelatha et al., 2026). **Benchmark suites** broaden prompt and attribute coverage further (scenes, objects, clothing, safety) and consistently

report stereotyping and amplification effects, with target fairness the norm and threshold fairness rare (Li et al., 2025b; Lee et al., 2023; D'Inca et al., 2024; Bakr et al., 2023; Luccioni et al., 2023; Jin et al., 2024). Across all three spaces, threshold fairness remains the exception rather than the rule. A full breakdown of findings by evaluation space, including the specific demographic/contextual attributes and bias types identified in each, is provided in Appendix B.

5. Mitigation Strategies

Researchers have proposed interventions at every stage of the T2I pipeline to counteract societal biases learned from training data. We categorize these strategies into four intervention points (visualized in Fig. 1): *Prompt Engineering*, *Model Realignment*, *Embedding Manipulation*, and *Diffusion Guidance*. **Prompt engineering** appends demographic descriptors to neutral prompts at inference time; it is lightweight and applicable to black-box models, but shifts the burden of fairness onto the user and leaves the underlying representational bias unaddressed (Bonna et al., 2024; Bansal, 2024; Cheng et al., 2024; Friedrich et al., 2025; Hao et al., 2023). **Model realignment** uses parameter-efficient fine-tuning (e.g., LoRA-based distribution alignment, concept editing) to steer or "unlearn" biased associations directly in the model weights, addressing bias closer to its root but at higher computational cost (Shen et al., 2023; Hu et al., 2022; Gandikota et al., 2024a,b). **Embedding manipulation** intervenes earlier, in the text encoder, via geometric projection, additive steering, or token-level optimization to decouple profession terms from demographic associations before they reach the generative backbone (Thirugnana Sambandham and Schön, 2026; Chuang et al., 2023; Li et al., 2025a; Hirota et al., 2025; Zhang et al., 2023; Kim et al., 2023, 2025; Vadakkeeveetil Sreelatha et al., 2026). **Diffusion guidance** steers the denoising trajectory dynamically at inference time, allowing precise attribute control without retraining, at the cost of added inference latency from per-step gradient calculations (Parihar et al., 2024; Kang et al., 2025; Friedrich et al., 2025; Hertz et al., 2022; Brack et al., 2023; Azam and Akhtar, 2025; Meng et al., 2021). A full account of methods within each category, including trade-offs and reproducibility gaps for individual techniques, is provided in Appendix C.

6. Discussion

In this section, we discuss tensions in how fairness is defined and applied, the normative justifications offered, and gaps in mitigation strategies identified in our survey. We close by proposing pragmatic standards.

6.1. Tension in fairness operationalization: target vs threshold fairness

In our survey of the T2I fairness/bias landscape, we found that fairness is operationalized in different ways across evaluation spaces. In our framework, we make a distinction between research that states a normative ideal (target fairness) and research that sets tolerance bands, uncertainty, and a pass/fail rule (threshold fairness). Both types of fairness have their use in T2I evaluation/mitigation spaces. Target fairness, for example, is better aligned with descriptive or exploratory studies, like diagnosis, model comparison, early benchmarks, and qualitative audits.

Threshold fairness, however, is warranted if fairness is being recognized as a claim (like certification, approval, deployability) or otherwise in works that claim fairness has been achieved. When such claims are made (most often in benchmarks and automated audits like

analysis, VQA, and caption-based evaluations), we would expect the use of threshold fairness so that reported disparities are judged against pre-registered acceptability criterion and are robust to sampling and measurement uncertainty. Making these claims without uncertainty or threshold rules ignores how much deviation from the target counts as "acceptable," given prompt/seed variability and imperfect attribute labeling. Ignoring this leaves space for the following issues: (1) results are hard to compare across papers because the acceptability gap is not explicit, and (2) mitigation claims can be overstated because any gap reduction is presented as fairness being achieved. As such, the absence of threshold fairness in these evaluation spaces risks leaving many entries into T2I research closer to descriptive audits, rather than enforcement of their stated normative fairness ideals.

6.2. Tension between declared and operationalized fairness

We also came across tensions in the naming vs the operationalization of fairness and bias. Some works claim to measure bias while actually implementing fairness, and vice versa. Naik and Nushi (2023) is cast as an analysis of social bias, though its results are interpreted against an external representational benchmark, which introduces a normative comparison more commonly associated with fairness evaluation. Sun et al. (2024) presents its analysis as one of occupational gender bias, its use of census and other reference distributions introduces a normative comparative baseline that also makes the evaluation partly fairness relevant. We note this tension because the separation of bias (descriptive disparities) from fairness (a defended ideal and a decision rule) would reduce nominative confusion, and make comparison between studies easier.

6.3. Fairness targets and the (lack of) normative defense

We note that across these papers, precise normative ideals (like a uniform gender distribution target) are largely undefended. We recognize that some metrics are easier to use in some evaluation spaces. However, it's important to note that normative targets presuppose a value claim, in which equal shares and/or matching to real-world statistics are inherently fair. This assumption ignores base rates, can conflict with other desirable objectives (calibration, error-rate parity), and is Goodhart-prone: i.e., you can hit 50/50 while still under-serving a group (Kleinberg et al., 2017; Manheim and Garrabrant, 2018). We contend that 50/50 demographic parity and proportional representation should not be treated as an unexplained universal ideal (Friedler et al., 2021). We propose that authors explain why they aim for a target, what normative trade-offs it implies, and how uncertainty and error behavior are handled. This is especially important as research is ongoing in terms of normative philosophy as well as user expectations of T2I outputs (Richter et al., 2025).

6.4. Critical Gaps in Bias Mitigation

Our review identifies a fundamental tension between efficacy and feasibility in current mitigation strategies. Inference-time interventions (e.g., prompt engineering) offer lightweight accessibility for black-box models (Luo et al., 2024; Hao et al., 2023). Conversely, deeper interventions like fine-tuning (Shen et al., 2023; Gandikota et al., 2024a) or embedding manipulation (Chuang et al., 2023; Li et al., 2025a; Zhang et al., 2023) address root causes but incur prohibitive computational costs and often lack holistic pipeline coverage. In addition to these concern, there is also a reproducibility crisis; code unavailability for key methods (Kim et al., 2025; Hirota et al., 2025) restricts rigorous comparison of these trade-offs.

Furthermore, the field lacks standardized mitigation benchmark. Unlike bias detection methods (Li et al., 2025b; D'Inca et al., 2024), mitigation claims often rely on ad-hoc metrics that obscure degradation in other areas. Further benchmarking efforts must simultaneously evaluate five competing axes: (1) Bias Reduction (against a defined target), (2) Image Fidelity, (3) Prompt Adherence, (4) Generalizability across unseen concepts, and (5) Negative Side Effects. Addressing these gaps require a community-driven shift toward open-source evaluation harnesses that can rigorously test complex, intersectional alignment failures beyond single-axis attributes.

7. Conclusion

In this paper we survey the use of fairness and bias across T2I works. To achieve this, we normalize definitions of fairness and bias, then identify patterns in the use of fairness, bias, and demographic/contextual attributes across evaluation and mitigation spaces. Based off of these findings, we propose a new taxonomy of fairness types which delineates normative-only fairness (target fairness) and decision rule, pass/fail based fairness (threshold fairness). We establish an appropriate range of use for both fairness types, while expanding on the importance of distinguishing between them. Additionally, we identify nominative tension between declared and operationalized fairness/bias types in some T2I works. We also note how normative justification is left implicit in some works, in addition to identifying gaps in bias mitigation. This survey categorizes evaluation spaces, mitigation techniques, and the use of attributes, bias, and fairness notions across T2I literature up until the time of writing. In doing so we argue for the normalization of fairness and bias terminology, and provide a new taxonomy towards this end. Full details of our literature search and screening methodology are provided in Appendix A.

Author Contributions

Megan Smith conducted the initial literature review on fairness notions and bias types. Venkatesh Thirugnana Sambandham conducted literature review on the foundations and mitigation strategies. Florian Richter, Matthias Uhl, Laura Crompton, and Torsten Schön supervised the project, assisted in organizing the study, and contributed to refining the manuscript.

Acknowledgments

This project is funded by the Bayerische Transformationsund Forschungsstiftung under the project EvenFAIr (AZ-1611-23).

References

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

Mikko Apiola, Henriikka Vartiainen, and Matti Tedre. First year cs students exploring and identifying biases and social injustices in text-to-image generative ai. pages 485–491, 07 2024. doi: 10.1145/3649217.3653596.

Basim Azam and Naveed Akhtar. Plug-and-play interpretable responsible text-to-image generation via dual-space multi-facet concept control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2976–2985, 2025.

Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20041–20053, 2023.

Prashant Bansal. Prompt engineering importance and applicability with generative ai. *Journal of Computer and Communications*, 12(10):14–23, 2024.

Kimberley Baxter. Uncovering Bias in AI-Generated Images. In *Proceedings of the ART of Safety: Workshop on Adversarial testing and Red-Teaming for generative AI*, pages 49–50, Bali, Indonesia, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.artofsafety-1.6. URL https://aclanthology.org/2023.artofsafety-1.6.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1493–1504, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594095. URL https://doi.org/10.1145/3593013.3594095.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Sarah Bonna, Yu-Cheng Huang, Ekaterina Novozhilova, Sejin Paik, Zhengyang Shan, Michelle Yilin Feng, Ge Gao, Yonish Tayal, Rushil Kulkarni, Jialin Yu, et al. Debi-aspi: inference-time debiasing by prompt iteration of a text-to-image generative model. In *European Conference on Computer Vision*, pages 68–83. Springer, 2024.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems*, 36:25365–25389, 2023.

Tianwei Chen, Yusuke Hirota, Mayu Otani, Noa Garcia, and Yuta Nakashima. Would deep generative models amplify bias in future models? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10833–10843, 2024.

Chih-Hong Cheng, Harald Ruess, Changshun Wu, and Xingyu Zhao. Conditional fairness for generative ais. *arXiv preprint arXiv:2404.16663*, 2024.

Chih-Hong Cheng, Changshun Wu, Xingyu Zhao, Saddek Bensalem, and Harald Ruess. Runtime monitoring and enforcement of conditional fairness in generative ais. In *International Conference on Runtime Verification*, pages 73–91. Springer, 2025.

Marc Cheong, Ehsan Abedin, Marinus Ferreira, Ritsaart Reimann, Shalom Chalson, Pamela Robinson, Joanne Byrne, Leah Ruppanner, Mark Alfano, and Colin Klein. Investigating gender and racial biases in dall-e mini images. In *ACM Responsible Computing*, 2024.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills social biases of text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

Sam Corbett-Davies, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312): 1–117, 2023.

Geoffrey Currie, George John, and Johnathan Hewis. Gender and ethnicity bias in generative artificial intelligence text-to-image depiction of pharmacists. *International Journal of Pharmacy Practice*, 32, 09 2024. doi: 10.1093/ijpp/riae049.

Moreno D'Inca, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Openbias: Open-set bias detection in text-to-image generative models. In *CVPR*, 2024.

Wala Elsharif, Mahmood Alzubaidi, and Marco Agus. Cultural bias in text-to-image models: A systematic review of bias identification, evaluation and mitigation strategies. *IEEE Access*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Antonio Ferrara, Francesco Cozzi, Alan Perotti, André Panisson, and Francesco Bonchi. Size-adaptive hypothesis testing for fairness, 2025. URL https://arxiv.org/abs/2506.10586.

Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143, 2021. doi: 10.1145/3433949.

Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Auditing and instructing text-to-image generation models on fairness. *AI and Ethics*, 5(3):2103–2123, 2025.

Zihao Fu, Ryan Brown, Shun Shao, Kai Rawal, Eoin Delaney, and Chris Russell. Fairimagen: Post-processing for bias mitigation in text-to-image models. *Advances in Neural Information Processing Systems*, 38:29488–29512, 2026.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pages 172–188. Springer, 2024a.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024b.

Eva Gengler. Sexism, racism, and classism: Social biases in text-to-image generative ai in the context of power, success, and beauty. In *Wirtschaftsinformatik 2024 Proceedings*, 11 2024.

Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson, and Aylin Caliskan. Do generative ai models output harm while representing non-western cultures: Evidence from a community-centered approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 476–489, 2024.

Sourojit Ghosh, Nina Lutz, and Aylin Caliskan. *"I Don't See Myself Represented Here at All": User Experiences of Stable Diffusion Outputs Containing Representational Harms across Gender Identities and Nationalities*, page 463–475. AAAI Press, 2025.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.

Douglas Guilbeault, Solène Delecourt, Tasker Hull, Bhargav Srinivasa Desikan, Mark Chu, and Ethan Nadler. Online images amplify gender bias. *Nature*, 626(8001):1049–1055, 2024.

Shahin Hakemi, Naveed Akhtar, Ghulam Mubashar Hassan, and Ajmal Mian. Deeper diffusion models amplify bias. *arXiv preprint arXiv:2505.17560*, 2025.

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. *Advances in Neural Information Processing Systems*, 36:63687–63723, 2023.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939, 2023.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528, 2021.

Yusuke Hirota, Min-Hung Chen, Chien-Yi Wang, Yuta Nakashima, Yu-Chiang Frank Wang, and Ryo Hachiuma. Saner: Annotation-free societal attribute neutralizer for debiasing clip. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023.

Di Jin, Xing Liu, Yu Liu, Jia Qing Yap, Andrea Wong, Adriana Crespo, Qi Lin, Zhiyuan Yin, Qiang Yan, and Ryan Ye. Infelm: In-depth fairness evaluation of large text-to-image models. *arXiv preprint arXiv:2501.01973*, 2024.

Sangwon Jung, Alex Oesterling, Claudio Mayrink Verdun, Sajani Vithana, Taesup Moon, and Flavio P Calmon. Multi-group proportional representations for text-to-image models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23744–23754, 2025.

Mintong Kang, Vinayshekhar Bannihatti Kumar, Shamik Roy, Abhishek Kumar, Sopan Khosla, Balakrishnan Murali Narayanaswamy, and Rashmi Gangadharaiah. Fairgen: Controlling sensitive attributes for fair generations in diffusion models via adaptive latent guidance. *arXiv preprint arXiv:2503.01872*, 2025.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, pages 3819–3828, 2015.

Eunji Kim, Siwon Kim, Chaehun Shin, and Sungroh Yoon. De-stereotyping text-to-image models through prompt tuning. 2023.

Eunji Kim, Siwon Kim, Minjun Park, Rahim Entezari, and Sungroh Yoon. Rethinking training for de-biasing text-to-image generation: Unlocking the potential of stable diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13361–13370, 2025.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017. doi: 10.4230/LIPIcs.ITCS.2017.43. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2017.43.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011, 2023.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12006–12016, 2024.

Jia Li, Lijie Hu, Jingfeng Zhang, Tianhang Zheng, Hua Zhang, and Di Wang. Fair text-to-image diffusion via fair mapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26256–26264, 2025a.

Lijun Li, Zhelun Shi, Xuhao Hu, Bowen Dong, Yiran Qin, Xihui Liu, Lu Sheng, and Jing Shao. T2isafety: Benchmark for assessing fairness, toxicity, and privacy in image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13381–13392, 2025b.

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

Hanjun Luo, Ziye Deng, Haoyu Huang, Xuecheng Liu, Ruizhe Chen, and Zuozhu 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.

Yunbo Lyu, Zhou Yang, Yuqing Niu, Jing Jiang, and David Lo. Do existing testing tools really uncover gender bias in text-to-image models? *arXiv preprint arXiv:2501.00000*, 2025.

Kelly Avery Mack, Rida Qadri, Remi Denton, Shaun K. Kane, and Cynthia L. Bennett. "they only care to show us the wheelchair": disability representation in text-to-image ai models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642166. URL https://doi.org/10.1145/3613904.3642166.

Sina Malakouti and Adriana Kovashka. Role bias in diffusion models: Diagnosing and mitigating through intermediate decomposition. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018. doi: 10.48550/arXiv.1803.04585. URL https://arxiv.org/abs/1803.04585.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 786–808, New York, NY, USA, 2023. Association for Computing Machinery.

ISBN 9798400702310. doi: 10.1145/3600211.3604711. URL https://doi.org/10.1145/3600211.3604711.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R Venkatesh Babu. Balancing act: distribution-guided debiasing in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6668–6678, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, October 2023.

Malsha V Perera and Vishal M Patel. Analyzing bias in diffusion-based face generation models. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023.

Rida Qadri, Renee Shelby, Cynthia Bennett, and Emily Denton. Ai's regimes of representation: A community-centered study of text-to-image models in south asia. pages 506–517, 06 2023. doi: 10.1145/3593013.3594016.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67, 2020.

Florian Richter, Sebastian Krügel, and Matthias Uhl. Users prefer generative ai to depict what is real instead of what is ideal. *SSRN*, 2025.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Teresa Sandoval-Martin and Ester Martínez-Sanzo. 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. ISSN 2076-0760. doi: 10.3390/socsci13050250. URL https://www.mdpi.com/2076-0760/13/5/250.

Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6367–6384, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.353. URL https://aclanthology.org/2024.naacl-long.353/.

Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Finetuning text-to-image diffusion models for fairness. *arXiv preprint arXiv:2311.07604*, 2023.

Silpa Vadakkeeveetil Sreelatha, Dan Wang, Serge Belongie, Muhammad Awais, and Anjan Dutta. Raigen: Rare attribute identification in text-to-image generative models. *arXiv preprint arXiv:2602.06806*, 2026.

Luhang Sun, Mian Wei, Yibing Sun, Yoo Ji Suh, Liwei Shen, and Sijia Yang. Smiling women pitching down: auditing representational and presentational gender biases in image-generative ai. *Journal of Computer-Mediated Communication*, 29(1):zmad045, 02 2024. ISSN 1083-6101. doi: 10.1093/jcmc/zmad045. URL https://doi.org/10.1093/jcmc/zmad045.

Piotr Szymański, Magdalena Lipczyńska, and Anna Maria Górska. From data to perception: visualizing bias in artificial intelligence-generated images. *European Heart Journal*, page ehae850, March 2025. ISSN 0195-668X, 1522-9645. doi: 10.1093/eurheartj/ehae850. URL https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehae850/8088193.

Christopher T Teo, Milad Abdollahzadeh, Xinda Ma, and Ngai-man Cheung. Fairqueue: Rethinking prompt learning for fair text-to-image generation. *Advances in Neural Information Processing Systems*, 37:22878–22926, 2024.

Venkatesh Thirugnana Sambandham and Torsten Schön. Embedding arithmetic: A lightweight, tuning-free framework for post-hoc bias mitigation in text-to-image models. In *International Conference on Pattern Recognition*, pages 308–323. Springer, 2026.

Silpa Vadakkeeveetil Sreelatha, Sauradip Nag, Muhammad Awais, Serge Belongie, and Anjan Dutta. Respodiff: Dual-module bottleneck transformation for responsible & faithful t2i generation. *Advances in Neural Information Processing Systems*, 38:42551–42583, 2026.

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

Jialu Wang, Xinyue Liu, Zonglin Di, Yang Liu, and Xin Wang. T2iat: Measuring valence and stereotypical biases in text-to-image generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2560–2574, 2023.

Yankun Wu, Yuta Nakashima, and Noa Garcia. Revealing Gender Bias from Prompt to Image in Stable Diffusion. *Journal of Imaging*, 11(2):35, January 2025. ISSN 2313-433X. doi: 10.3390/jimaging11020035. URL https://www.mdpi.com/2313-433X/11/2/35.

Eric J. York, Eva Brumberger, and La Verne Abe Harris. Prompting bias: Assessing representation and accuracy in ai-generated images. In *Proceedings of the 42nd ACM International Conference on Design of Communication*, SIGDOC '24, page 106–115, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705199. doi: 10.1145/3641237.3691658. URL https://doi.org/10.1145/3641237.3691658.

Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Iti-gen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3969–3980, October 2023.

Yifeng Zhang, Ming Jiang, and Qi Zhao. Grace: graph-based contextual debiasing for fair visual question answering. In *European Conference on Computer Vision*, pages 176–194. Springer, 2024.