IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 April 2025, accepted 28 June 2025, date of publication 3 July 2025, date of current version 18 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3585745

TOPICAL REVIEW

Cultural Bias in Text-to-Image Models: A Systematic Review of Bias Identification, Evaluation, and Mitigation Strategies

WALA ELSHARIF, MAHMOOD ALZUBAIDI, AND MARCO AGUS, (Member, IEEE)
ICT Division, College of Science and Engineering, Hamad Bin Khalifa University, Ar Rayyan, Qatar

Corresponding author: Marco Agus (magus@hbku.edu.qa)

Open Access funding provided by the Qatar National Library.

ABSTRACT Despite their continuous advancements, text-to-image (TTI) models often reflect and reinforce cultural biases, perpetuating stereotypes often inherent in their training data. This systematic review critically examines cultural bias in text-to-image (TTI) models, addressing gaps in existing research by analyzing its manifestations, evaluation methods, and mitigation strategies—both directly and through the lens of intersectionality with other bias dimensions. A comprehensive literature review was conducted across multiple major databases, following a rigorously structured search strategy, resulting in the selection of 58 studies spanning bias analysis, evaluation frameworks, and mitigation techniques. Thematic findings highlight that gender bias was the most extensively studied, appearing in 53 studies (91%), followed by racial/ethnic bias (42 studies) and other social biases (41 studies). Furthermore, the review explores how these biases intersect and compound in AI-generated imagery, shaping and reinforcing cultural bias. Our findings reveal the following key aspects: 1) the lack of standardization and scalability in bias evaluation, 2) the lack of a fully effective mitigation strategy, 3) contributed TTI benchmarks favoring Western-centric perspectives. We finally propose future directions to improve fairness and representation in TTI models.

INDEX TERMS AI ethics, AI fairness, bias evaluation, bias mitigation, CLIP, cultural bias, generative AI, gender bias, prompt engineering, racial bias, responsible AI, text-to-image models.

I. INTRODUCTION

Text-to-image (TTI) generation [1] is a rapidly evolving field in artificial intelligence (AI) that focuses on creating visual content from textual descriptions. This technology has found applications in diverse areas such as creative arts, virtual reality, assistive tools for the visually impaired, education, and entertainment [2], [3], [4], [5], [6].

Recent advances in deep learning, particularly in generative models, have significantly improved the quality, resolution, and semantic alignment of generated images [7], [8]. Moreover, Contrastive Language-Image Pretraining (CLIP) and CLIP-based TTI models have enhanced multimodal AI by aligning textual descriptions with visual representations [9]. CLIP, introduced in 2021, enabled zero-shot image classification, laying the foundation for generative models such as DALL-E and Stable Diffusion [1], [10]. These innovations, combined with the growing accessibility of TTI tools, have transformed how AI systems interpret and visualize human language.

However, alongside their benefits, these models carry inherent biases reflective of the datasets they are trained on and the socio-cultural contexts from which these datasets emerge [11], [12]. Cultural bias in particular has emerged as a key concern, encompassing dimensions such as gender, race, and class, and often manifesting through stereotypes or the exclusion of underrepresented groups. For example, stereotypical depictions of male and female professions frequently occur. Figure 1 illustrates the biases present in TTI generation by showcasing images produced using the

The associate editor coordinating the review of this manuscript and approving it for publication was Syed Islam.

TABLE 1. Comparison of previous relevant reviews on TTI bias to our systematic review.

Study	Study Focus		Bias Identification and Measurement		Bias Mitigation	
	Bias Dimensions	TTI Models	Identification Methods	Measurement Metrics	Methods Review	Methods Effectiveness
Bird et al. [13]. (2023)	Harmful Associations	✓	✓	×	Partial	
Parraga et al. [14]. (2023)	Fairness in deep learning	Partial	✓	✓	✓	✓
Nemani et al. [15]. (2024)	Gender	Partial	✓	✓	✓	✓
Prerak [16]. (2024)	Gender, skin tone, and geo-cultural	Partial	Partial	×	Partial	×
Saxena [17]	Partial	×	✓	✓	×	×
Ours	Cultural (intersectionality of gender, racial and social biases)	✓	✓	✓	✓	✓

FIGURE 1. Sample of AI-generated images, generated using Stable Diffusion with the prompt "A doctor" on the first row and "A nurse" in the second row. The images illustrate a clear gendered professional stereotype, with 'doctor' exclusively associated with men and 'nurse' with women—reflecting Western-centric cultural norms. Additionally, across eight samples, all generated individuals appear white, highlighting racial bias. These consistent patterns suggest that even neutral prompts activate entrenched demographic assumptions in Stable Diffusion XL.

FIGURE 2. Sample of AI-generated images, generated using Stable Diffusion with the prompt "A Middle Eastern teacher" on the first row, the images showcase gender and racial/ethnic bias as it limits the results to male teachers reflecting that middle eastern teachers are exclusively men. In the second row, the neutral prompt "A teacher" is used, also portraying the same biases as a white female is consistently depicted, reflectional the Western conception that teaching jobs are often occupied by females.

prompts "A doctor" and "A nurse" with Stable Diffusion XL. The generated images possibly reveal racial bias, where a race-neutral prompt predominantly depicts individuals of white ethnicity. Furthermore, the outputs reinforce western gender stereotypes by consistently portraying doctors as male and nurses as female.

Similarly, racial bias manifests in the erasure or misrepresentation of non-Western ethnicities, contributing to the marginalization of diverse racial identities or portraying misconceptions of certain races [18]. These biases are often compounded by other factors such as age and class, forming intersecting layers of social inequality. Cultural differences also play a significant role—what is considered a stereotype in one cultural context may be interpreted differently in another.

While prior research has primarily focused on gender and racial bias in TTI models, cultural bias remains underexplored. Common approaches to bias detection include embedding association tests and fairness metrics, while mitigation efforts have ranged from prompt engineering to dataset curation. Bird et al. [13] examine multiple harmful associations in TTI models, including cultural bias, yet their review lacks a comprehensive analysis of identification and mitigation methods. Parraga et al. [14] provide a systematic review of fairness in AI models, including TTI, but do not focus on cultural bias or TTI models specifically. Nemani et al. [15] investigate gender bias in transformer models, covering identification and mitigation methods, though their survey is limited to gender bias and does not address TTI. Saxena [17] conduct a comprehensive narrative review on quantitative measures of bias across AI-generated modalities but do not explore identification or mitigation strategies. Prerak [16] review gender, skin tone, and geo-cultural biases in TTI models, covering only a subset of evaluation and mitigation methods, omitting several key metrics and a detailed assessment of mitigation strategies and their effectiveness.

This review aims to fill a crucial gap by focusing explicitly on cultural bias in TTI models and its intersection with other forms of bias. While prior work has predominantly

FIGURE 3. Sample of AI-generated images, generated using Stable Diffusion XL with the prompt "A poor person" on the first row, and "A rich person" in the second row. The images showcase an intersection of gender bias and racism. The model consistently depicts a rich person as black females whereas a rich person is a white male.

FIGURE 4. Sample of AI-generated images, generated using Stable Diffusion with the prompt "An American family" on the first row and "An Asian family" in the second row. The images showcase racial or ethnic bias where the model only depicts a certain race or ethnicity for an attribute that bears multiple races.

addressed gender and racial/ethnic biases in isolation, the cultural dimension—and how it compounds with these social biases remains underexplored. Table 1 summarizes our contributions in contrast to prior reviews. This review conducts a systematic examination of the literature covering cultural bias and intersecting biases in CLIP-based TTI models, including gender and racial/ethnic stereotypes that contribute to cultural misrepresentation. We first analyze how different types of bias are identified, assessed, and measured in generated images. We then review approaches for bias mitigation, including interventions in model architecture, dataset composition, and prompt engineering strategies.

Our contributions in this work are summarized as:

1) A systematic review of existing literature on cultural bias in CLIP-based TTI models in works that address it directly or through its compounding dimensions: racial/ethnic, gender and social biases.
2) An in-depth review and analysis of identification, evaluation and measurement of the investigated biases.
3) A comparative analysis of bias mitigation methods in TTI generation and their effectiveness.
4) Identification of research gaps and actionable recommendations to mitigate cultural bias in TTI models, fostering inclusivity and diversity in AI-generated imagery.

The rest of this review first provides a background on TTI generation and cultural bias in TTI models in Section II. followed by detailed methodology of the review in Section III. We then show the results in detail, explaining the selected studies in the review and answering the research question followed by thematic findings of the research in Section IV. A discussion section is then provided in Section VI and the limitations of the research are explained, and finally, the systematic review and its findings are concluded in Section VII.

II. BACKGROUND AND RELATED WORK

This section provides an overview of the main components of this systematic review. First, it delves into the key technologies, models, and recent advances that have shaped the field of TTI generation. Next, an overview is provided on cultural and the dimensions that compose it. Finally, it discusses the methods of bias evaluation and mitigation that have been developed.

A. OVERVIEW OF TTI MODELS

At the core of TTI generation is the fusion of natural language processing (NLP), computer vision, and generative modeling. NLP is pivotal in interpreting and encoding textual descriptions into a format suitable for generative models. Language models that have been pre-trained, like BERT, GPT, and T5, are typically used to derive semantic features from the text [32], [33], [34]. These features are then converted into embeddings that steer the image creation process. By capturing the subtleties of language, NLP ensures that the produced images match closely the input text.

Generative Adversarial Networks (GANs) have been a fundamental component of TTI generation since their introduction [35]. A GAN consists of two neural networks: a generator and a discriminator. The generator creates images from textual embeddings, while the discriminator evaluates the realism and quality of these images. Through an adversarial process, the generator learns to produce increasingly realistic and semantically aligned images. Models like AttnGAN [36] and StackGAN [37] have demonstrated the ability to generate high-resolution images by refining the generation process in multiple stages. These models incorporate attention mechanisms to focus on specific parts of the text, ensuring finer details in the generated images.

Another important class of generative models used in TTI generation is Variational Autoencoders (VAEs) [38]. VAEs encode input data into a latent space and then decode it to generate new samples. In the context of TTI generation, VAEs are often combined with text encoders to map textual descriptions into the latent space, which is then decoded into images. While VAEs tend to produce smoother and more diverse outputs compared to GANs, they often struggle with generating high-resolution images. Hybrid approaches, such as VAE-GANs, have been proposed to combine the strengths of both models, achieving a balance between image quality and diversity [39].

TABLE 2. Summary of CLIP and popular state-of-the art CLIP-based text-to-image models.

Model	Release Year	Training Dataset	Main Contribution
CLIP [9]	2021	LAION-400M [19], WIT-400M [20]	Zero-shot image classification with vision-language alignment
DALL·E [1]	2021	250M image-text pairs (Web-scraped)	Autoregressive transformer for text-to-image generation with CLIP evaluation
GLIDE	2021	Web-scraped image-text pairs	Diffusion models for text-to-image generation with improved quality over autoregressive methods
CLIP-Guided VQGAN	2021	ImageNet [21] + Web images	Fine-grained text-based image manipulation using CLIP embeddings and GANs
BigGAN + CLIP [22]	2021	ImageNet	CLIP-guided BigGAN generations for improved text alignment
DALL-E 2 [23]	2022	CLIP (prior-based)	Improved image quality and text-image coherence
Stable Diffusion v1 [10]	2022	LAION-5B [24]	Open-source, efficient text-to-image generation
Stable Diffusion v2 [25]	2022	LAION-5B	Enhanced image quality, better negative prompting
Stable Diffusion XL [26]	2023	LAION-5B	Higher resolution images, better text understanding
Imagen [7]	2022	LAION-400M, CC3M [27], CC12M [28]	High-quality image generation via diffusion models
Parti [29]	2022	LAION-400M	Autoregressive transformers for photorealistic image generation
DeepFloyd IF [30]	2023	LAION-5B	Cascading diffusion models for high-resolution photorealistic generation
DALL·E 3 [31]	2023	Filtered Web dataset	Enhanced prompt comprehension and output consistency with CLIP-based refinements

CLIP and CLIP-based TTI Models: CLIP is a multimodal model developed by OpenAI that learns to associate images and text by training on a large dataset of image-caption pairs [9]. It excels in zero-shot learning, enabling it to understand and generate visual concepts based on textual descriptions. Many state-of-the art TTI models incorporate CLIP, either fully or partially to enhance text-image alignment, refine outputs, and evaluate biases.

Although DALL·E 1 does not integrate CLIP in its generation process, it was later evaluated using CLIP to measure text-image consistency and rank outputs based on alignment [1]. DALL·E 2 builds on this by integrating CLIP's text and image encoders, mapping textual descriptions into a shared embedding space to improve coherence between prompts and images [23]. Stable Diffusion uses CLIP's text encoder to convert prompts into embeddings that guide the diffusion model [10].

GLIDE and DeepFloyd IF leverage CLIP to rank and refine generated images, ensuring closer alignment with textual descriptions [30], [40]. Other models, such as BigGANs and CLIP-Guided VQGAN, use CLIP's multi-modal embeddings for text-driven image modification [22], [41]. Meanwhile, DALL·E 3 builds on CLIP's architecture to enhance prompt comprehension and output consistency [31].

Despite these advancements, challenges remain in TTI generation. Ensuring fine-grained control over image attributes, improving the diversity of generated outputs, and addressing ethical concerns such as bias and misuse are ongoing areas of research. These models inherit biases from their training datasets, which are often web-scraped and reflect societal stereotypes. Bias in TTI models manifests in various forms, including gendered representations, cultural stereotypes, and over-representation of dominant groups, leading to fairness concerns in AI-generated imagery.

B. CULTURAL BIAS IN TTI MODELS

At its core, bias is defined as the lack of internal validity or incorrect assessment of the association between an exposure and an effect in the target population in which the statistic estimated has an expectation that does not equal the true value [21]. In a simple sense, bias is an inclination or prejudice for or against one person or group, especially in a way considered to be unfair. Hence, cultural bias refers to the tendency to interpret or judge phenomena by standards inherent to one's own culture [42].

Unlike biases that arise solely from demographic attributes such as gender or race, cultural bias is rooted in the norms, values, beliefs, traditions, and symbols of a given culture. It reflects how models trained on culturally dominant data may normalize certain worldviews, aesthetics, or practices while marginalizing or misrepresenting others. Cultural biases reflect differences between individuals and groups, often manifesting in varying preferences for specific characteristics. These differences can be observed across socio-economic status, language, race, ethnicity, and sexuality.

While cultural bias can intersect with other forms of social bias, such as racial or gender bias, it also operates independently—for example, in how models depict culturally-specific attire, religious rituals, or family

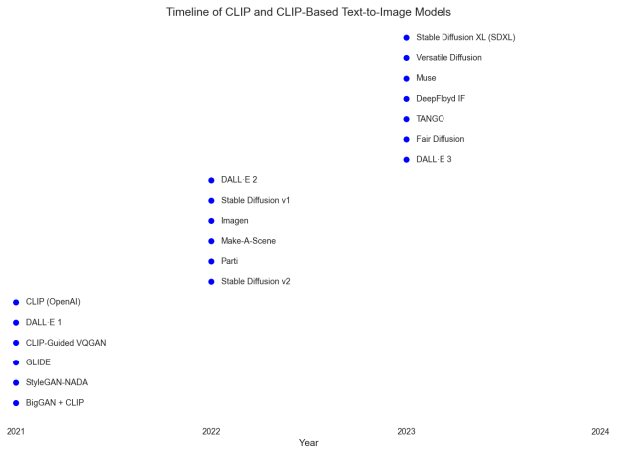

FIGURE 5. Timeline summarizing the development of CLIP and major state-of-the-art text-to-image (TTI) models alongside their respective training datasets. The figure illustrates the chronological progression of models from 2021 to 2024, marking a shift from early CLIP-guided models to more sophisticated diffusion-based architectures. It also highlights the increasing reliance on large-scale and multimodal datasets, underscoring the influence of dataset composition on the representational behavior and bias tendencies of each model generation.

structures. Cultural bias can reinforce stereotypes or misconceptions about certain cultures and, in some cases, contribute to racial and ethnic profiling [43].

Bias in TTI generation is a critical issue that reflects and amplifies societal inequalities, particularly in the areas of gender, race, and socio-economics. These biases often arise from the datasets used to train generative models, which may contain imbalanced or stereotypical representations of different groups [11]. In the context of TTI models, bias manifests in multiple dimensions, where certain groups are overrepresented while others are marginalized. This systematic review focuses on cultural bias as the main theme, exploring how different biases—gender, racial or ethnic, and class-based biases—interact with cultural norms and perspectives.

1) GENDER BIAS

Gender bias is the preference or prejudice toward one gender over the other [44]. It is frequently observed when models associate certain professions with specific genders, such as generating images of women for prompts like ''A nurse'' and men for prompts like ''A doctor'', as illustrated in Fig.1, a classic form of stereotype in Western culture. These outputs mirror historical and cultural stereotypes that have long defined gender roles in society. Such biases not only perpetuate outdated norms but also limit the representation of diverse gender identities in generated images. However, these biases are not uniform across cultures. In Western cultures, AI-generated images often align with traditional professional gender norms, reinforcing existing disparities in workforce representation. In contrast, some non-Western societies may have different cultural norms around gender roles that TTI models fail to capture due to predominantly Western-centric training data.

2) RACIAL/ETHNIC BIAS

Racial bias can be defined as a distortion arising from systemic, institutional, interpersonal or individual forms of explicit (conscious) or implicit (unconscious) prejudice against individuals or groups based on social constructs of race or ethnicity that influences the planning, methods, results, interpretation, dissemination and application of health research [45]. Another significant challenge in TTI generation, often resulting from datasets that lack diversity or overrepresent certain racial groups. For example, prompts like ''surgeon'' may predominantly generate images of lighter-skinned individuals [46], while prompts like ''athlete'' or ''manual laborer'' may overrepresent darker-skinned individuals. These biases stem from historical and cultural inequalities that have marginalized certain racial groups in professional and societal contexts. The consequences of such biases are far-reaching, as they reinforce harmful stereotypes and contribute to the underrepresentation of minority groups in positions of authority or prestige.

3) SOCIAL BIAS

Social bias refers to the unfair treatment or discrimination for or against an individual, group, or set of beliefs in a way that is prejudicial or unjust [47]. This form of bias is broad in scope, as it encompasses multiple dimensions by definition. It can appear in various ways, including biases related to age, occupation, ability, socioeconomic status, and other social factors. For example, older individuals may be underrepresented in tech-related imagery, reinforcing the stereotype that technology is primarily for younger generations [48]. Similarly, AI-generated images often depict high-status professions, such as CEOs or scientists, as predominantly male, while caregiving roles, such as teachers or nurses, are commonly portrayed as female. Religious and class-based biases can also emerge, with generative models frequently associating certain social classes with specific occupations or depicting religious attire in ways that reinforce stereotypes rather than diverse representations.

4) BIASES INTERSECTIONALITY

In 1989, Crenshaw and Bartlett [49] highlighted how multiple forms of discrimination—such as gender, race, and class—overlap, creating compounded disadvantages. In the context of TTI models, these overlapping biases do not merely coexist—they interact in ways that construct and reinforce culturally biased narratives. Intersectionality is thus not just a measure of compounded demographic bias but a key mechanism through which cultural misrepresentation is generated.

For instance, gender bias may lead to nurses being depicted as female, while racial bias reinforces that they are often women of color, whereas doctors are more likely to be white men as in Fig.1. Similarly, wealth is predominantly associated with white men, while poverty is assigned to racially

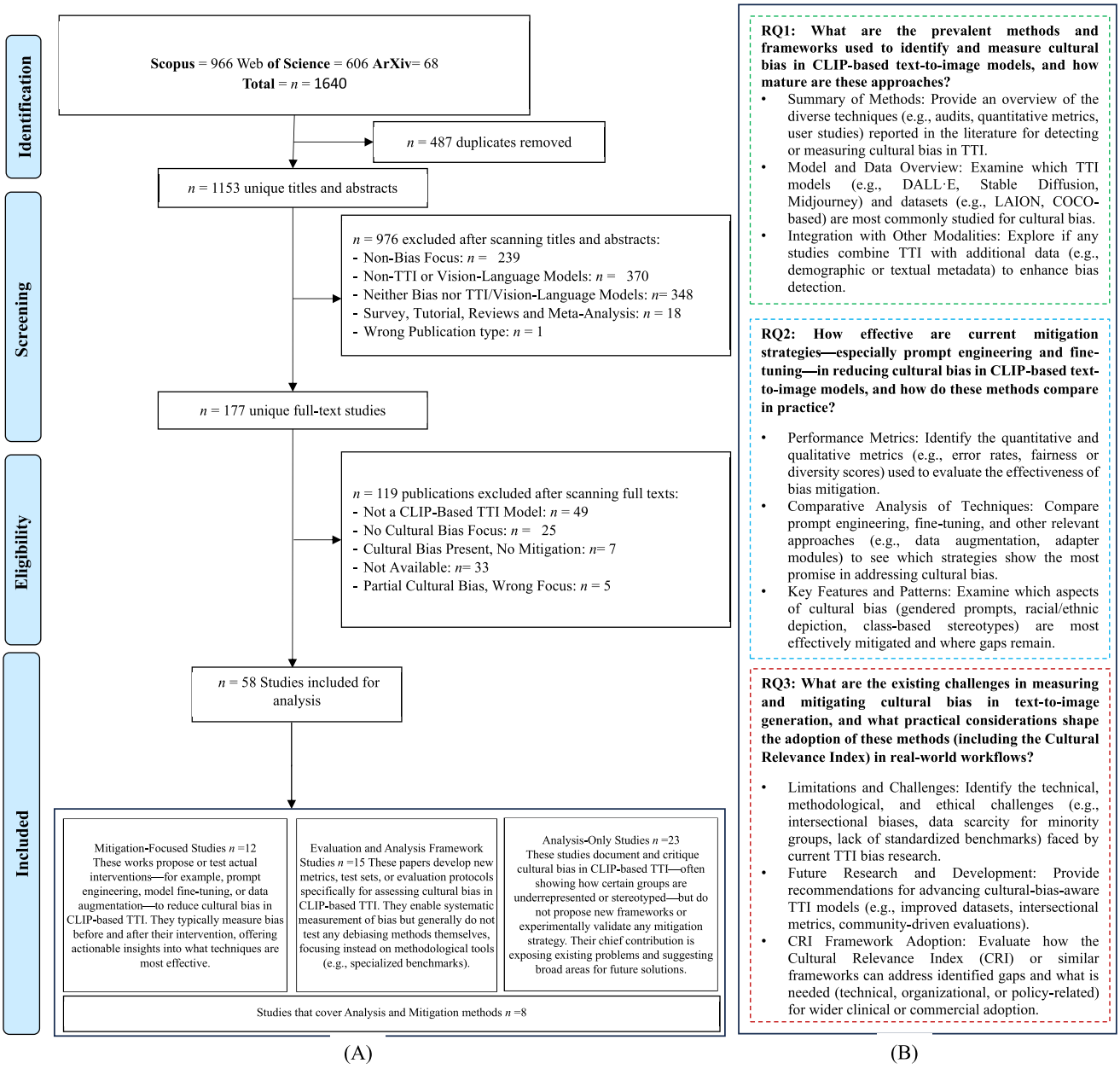

FIGURE 6. PRISMA flow diagram of study selection.

marginalized women, as in Fig.3, reinforcing entrenched cultural stereotypes of success and hardship.

These visual biases reflect more than statistical imbalance—they project dominant cultural assumptions about roles, status, and identity. Socioeconomic status intersects further, with high-paying professions often portrayed as white and male, while lower-wage jobs depict marginalized groups. Cultural bias also shapes representation, such as the monolithic portrayal of Muslim women solely in hijabs. Fig.2 shows the images generated with Stable Diffusion when prompted with "A Middle Eastern teacher", in the top row. The model tends to portray the teacher as a male while in reality, female teachers represent around half of the teaching force in the Middle East and North Africa Region [50]. On the other hand, in the second row, the model portrays the neutral prompt "A teacher" strictly as females, reflecting the Western bias of assigning caring occupations such as teaching to females.

However, the model still portrays biases that might stem from an inherent model misconception about the culture. Fig.4, on the other hand, showcases intersection of geo-cultural bias and racial bias, where prompts containing geographic terms like "American family" or "Asian family" repeatedly produce images featuring

a single race, ignoring racial diversity within those regions.

These intersecting biases work collectively to define and constrain cultural representations in generative outputs. They shape what is considered ''normal,'' ''professional,'' or ''traditional,'' based on dominant cultural lenses—often Western, urban, and affluent—while alternative cultural perspectives are rendered invisible or stereotyped. In this way, cultural bias is not merely the sum of demographic misrepresentations, but the outcome of how intersecting social hierarchies are encoded and reproduced visually by AI systems.

Addressing intersectional biases requires diverse training datasets and fairness metrics that account for overlapping cultural factors. Without such efforts, AI-generated imagery will continue to perpetuate systemic inequalities rather than fostering inclusive and accurate representations.

In addition, cross-cultural research highlights that biases in generative models often stem from cultural cues that shape perception and representation. For instance, construct bias occurs when culturally grounded concepts—like filial piety in East Asia versus the West—are represented differently, leading to incomplete or skewed visual interpretations. Similarly, stereotypes embedded in cultural narratives can reinforce reductive or exoticized depictions; as seen in Dagestani cultural contexts, where traits like ''hot-tempered'' or ''cunning'' are unfairly generalized across ethnic lines. These cultural cues, when internalized by models through training data, result in outputs that mirror entrenched social biases, often amplifying them. Therefore, biases in TTI systems are not just technical flaws—they are reflections of broader societal stereotypes and historical asymmetries that vary across cultural contexts [51], [52].

C. BIAS EVALUATION AND MITIGATION IN TTI MODELS

Bias in TTI models is typically identified and quantified through a combination of qualitative and quantitative approaches. Common evaluation methodologies include prompt-based analysis, where carefully crafted prompts are used to elicit and analyze biases in generated images [53], [54]. Another approach is embedding association tests (EATs), which extend techniques like the Word Embedding Association Test (WEAT) to measure implicit biases in multimodal systems [55]. Additionally, human evaluation remains a key method, as annotators can identify nuanced biases that automated tools may overlook [56], [57]. To scale analysis, researchers also leverage automated metrics and tools, such as CLIP-based models and vision-question answering (VQA) systems, to systematically quantify bias across large datasets [58], [59]. More recently, latent space analysis has provided insights into how biases are embedded within model architectures, offering new avenues for both detection and mitigation [60].

Mitigation strategies for bias in TTI models generally fall into four broad categories. Linguistic interventions modify prompts to reduce bias, though their effectiveness is often inconsistent [61], [62]. Model-centric adjustments, such as fine-tuning and architectural modifications, have demonstrated promise in reducing bias at a deeper level, with techniques like cross-attention disentanglement significantly improving fairness [63], [64]. Post-hoc corrections, including image filtering and debiasing frameworks, attempt to address biases in the final output, though they may not fully mitigate underlying model biases [65], [66]. Lastly, cross-modal approaches align text and image representations more equitably, leveraging multimodal learning techniques to reduce representational disparities [67], [68].

Despite advancements in bias evaluation and mitigation, challenges persist, including inconsistencies in bias measurement, scalability issues, and the evolving nature of cultural representations. Standardized benchmarks, such as BIGbench and FAIntbench, have improved bias quantification [58], [69], but issues like overcorrection and adversarial vulnerabilities remain. Addressing these challenges requires interdisciplinary collaboration and the development of robust, context-aware methodologies to ensure that generative AI models promote fair and inclusive representations rather than reinforcing existing stereotypes.

Key Findings and Research Gaps: Recent research has significantly advanced TTI models by integrating powerful architectures such as CLIP, diffusion models, and transformer-based encoders, which have improved image-text alignment and output fidelity. However, cultural representation remains a core challenge. While technical innovations enhance generation quality, they often perpetuate biases embedded in the underlying datasets. A growing body of work now investigates intersectional bias—gender, race, class, and culture—as critical to understanding fairness in generative outputs. Evaluative frameworks have become increasingly sophisticated, combining human annotation with large-scale automated analysis, yet consistency in metrics and cultural nuance detection remains limited. Mitigation strategies have evolved from prompt engineering to structural model changes and post-hoc corrections, but none offer comprehensive solutions. Notably, cross-modal and latent space interventions show emerging promise in addressing deep-seated representational biases. Despite this progress, key gaps persist, including the lack of culturally diverse training datasets, standardized bias benchmarks sensitive to cultural variation, and interdisciplinary approaches that bridge technical and socio-cultural perspectives.

III. METHODOLOGY

A. RESEARCH QUESTION

This systematic review investigates cultural bias in TTI models, particularly CLIP-based vision-language models. The key research questions guiding our review are:

- **RQ1:** What are the prevalent methods and frameworks used to identify and measure cultural bias in CLIP-based TTI models?

- **RQ2:** How effective are existing bias mitigation strategies—such as prompt engineering and fine-tuning—at reducing cultural bias in CLIP-based TTI models?
- **RQ3:** What challenges exist in measuring and mitigating cultural bias in TTI models, and what practical considerations influence their adoption in real-world applications?

B. SEARCH STRATEGY

A systematic search was conducted in five major databases: **Scopus, Web of Science, ArXiv** and **Google Scholar**. The search terms were designed to capture three key dimensions:

- **TTI Models:** Keywords included "text-to-image generation," "multimodal generative model," "CLIP model," "vision-language model," "diffusion model," "DALL-E," and "Stable Diffusion."
- **Bias Concepts:** Terms such as "bias," "prejudice," "stereotype," "unfairness," "disparity," "misrepresentation," and "cultural bias" were incorporated.
- **Evaluation and Mitigation Strategies:** Keywords included "bias measurement," "bias mitigation," "fair representation," "ethical AI", and "visual fairness."

Studies published from **2000 to 2024** were considered. The PRISMA framework was used to ensure a rigorous and transparent selection process.

C. INCLUSION AND EXCLUSION CRITERIA

Inclusion Criteria:

- Studies analyzing cultural bias in TTI models.
- Research investigating CLIP-based models, DALL-E, Stable Diffusion, or similar generative models.
- Papers evaluating or developing bias mitigation techniques (e.g., prompt engineering, fine-tuning, dataset interventions).
- Peer-reviewed journal articles or conference proceedings.
- Full-text availability in English.

Exclusion Criteria:

- Studies unrelated to cultural bias or fairness in AI.
- Research that does not involve TTI models.
- Reviews, meta-analyses, and tutorials without experimental results.
- Non-English publications or those with inaccessible full texts.

D. DATA EXTRACTION

Key metadata and analytical components were extracted from each included study:

- **Study metadata:** Title, authors, publication year, venue.
- **Research focus:** Core aim, hypothesis, and contribution.
- **Models Investigated:** CLIP, Stable Diffusion, DALL-E, Midjourney, etc.
- **Bias Category:** Gender, racial, ethnic, cultural, or societal biases.

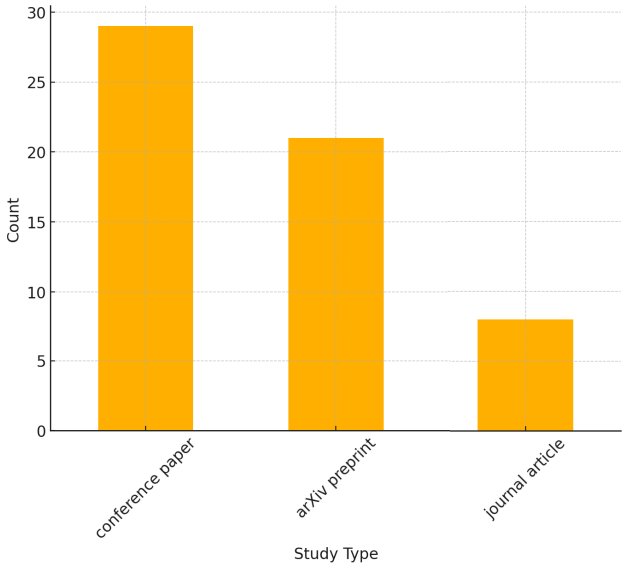

FIGURE 7. Bar graph illustrating the selected studies count per publication type: journal articles, conference papers or arXiv pre-prints. The chart reveals that conference papers constitute the largest share of the selected literature, followed by arXiv preprints and a smaller portion of peer-reviewed journal articles. This reflects the rapidly evolving nature of research in TTI bias, where timely dissemination often occurs via preprints and conferences prior to formal journal publication.

- **Mitigation Strategies:** Prompt engineering, fine-tuning, dataset augmentation.
- **Evaluation Metrics:** Fairness scores, diversity indices, subjective human assessments.
- **Outcomes:** Effectiveness of bias mitigation techniques, identified challenges.

All studies were coded for:

- **Bias type addressed** (e.g., gender, racial, ethnic).
- **Mitigation technique used** (e.g., prompting strategies, model adjustments).
- **Evaluation methods** (e.g., fairness scores, bias audits, expert reviews).

E. SYNTHESIS APPROACH

A combination of quantitative and qualitative methods was used to synthesize findings:

- **Descriptive Statistics:** Summarized publication trends, bias types, and model categories.
- **Comparative Evaluation:** Compared bias mitigation techniques based on reported effectiveness.
- **Thematic Analysis:** Identified common limitations, challenges, and best practices.
- **Visualization:** Findings were presented using tables, heatmaps, and radar charts for bias mitigation performance comparison.

IV. RESULTS

A. STUDY SELECTION OVERVIEW

A total of **1,640** studies were initially retrieved from three databases: Scopus ($n = 966$), Web of Science ($n = 606$), and

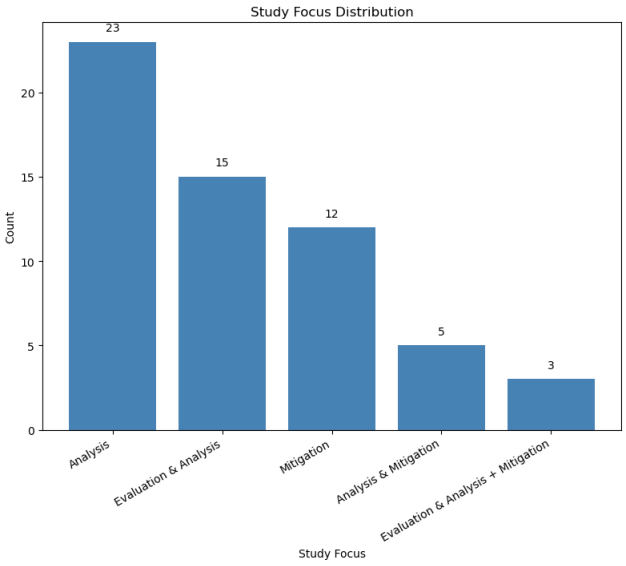

FIGURE 8. Bar graph illustrating the selected studies count per study focus. The figure shows that the majority of the studies focus on analysis-only approaches, indicating a research emphasis on identifying and describing bias over actively addressing it. A smaller portion of studies engage in both evaluation and mitigation, suggesting an underexplored area in bias reduction efforts. This imbalance underscores the need for more holistic frameworks that integrate analysis, evaluation, and mitigation strategies in TTI research. (Figure redrawn).

TABLE 3. Distribution of selected study types.

conference paper	arXiv preprint	journal article	Total
29	21	8	58

ArXiv ($n = 68$). After duplicate removal ($n = 487$), **1,153** unique studies remained for screening.

Following title and abstract screening, **976 studies** were excluded based on the following criteria:

- **Not focused on bias:** ($n = 239$)
- **Not a TTI or Vision-Language Model:** ($n = 370$)
- **Neither bias nor TTI/vision-language related:** ($n = 348$)
- **Survey, tutorial, or meta-analysis:** ($n = 18$)
- **Wrong publication type:** ($n = 1$)

After full-text screening ($n = 177$), an additional **118 studies** were excluded:

- **Not a CLIP-based TTI model:** ($n = 49$)
- **No cultural bias focus:** ($n = 25$)
- **Cultural bias present but no mitigation strategy:** ($n = 7$)
- **Full text not available:** ($n = 33$)
- **Partial cultural bias focus with incorrect emphasis:** ($n = 5$)

This resulted in **58 studies** included for final analysis. The study selection process is summarized in **Fig. 6**.

B. DESCRIPTIVE ANALYSIS OF INCLUDED STUDIES

The bar chart in Fig.9 illustrates the bias dimensions in the included studies in our review. It highlights the extensive research efforts that have been devoted to studying gender and racial/ethnic bias, social bias and its different manifestations have also been investigated in CLIP-based TTI models with varying.

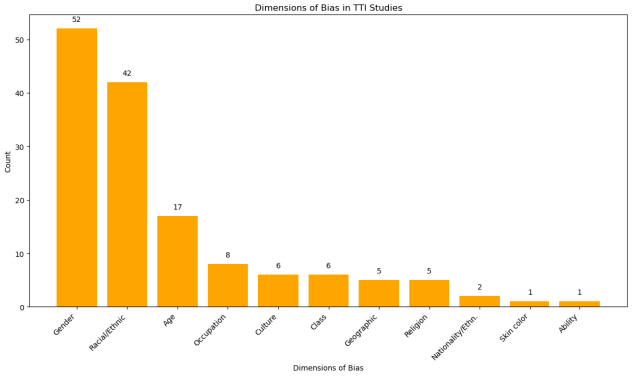

FIGURE 9. Bar chart of bias dimensions investigated in selected studies. The chart shows that gender and racial/ethnic bias are the two most investigated bias dimensions. The figure also reveals that other critical dimensions—such as culture, class, and religion—remain significantly underexplored. This disparity highlights a research gap where culturally nuanced biases and intersectional factors are often overlooked, despite their crucial role in shaping fair and inclusive TTI outputs. (Figure redrawn).

C. SYSTEMATIC MAPPING OF STUDY RESULTS

In this section, we map the studies included in the survey to answer the research question identified previously.

1) BIAS IDENTIFICATION AND MEASUREMENT

This subsection reviews methodologies employed to identify and measure biases in TTI models, categorizing 45 studies into *Evaluation and Analysis* (developing new metrics and protocols) and *Analysis Only* (documenting and critiquing biases). The studies are organized by methodology, highlighting key tools and findings.

a: PROMPT-BASED ANALYSIS

Prompt-based analysis is one of the most common methodologies for evaluating bias in TTI models. Researchers design prompts to elicit specific biases and analyze the generated images for demographic or thematic patterns. For example, Mannering [53] used paired male/female prompts to reveal gender associations in Stable Diffusion and DALL·E mini, finding that male prompts were associated with items like ties and trucks, while female prompts generated handbags and bowls. Similarly, Masrourisaadat et al. [54] employed 176 carefully designed prompts (88 for gender and 88 for racial bias) to generate 2,816 images, which were then categorized by human evaluators. Other studies, such as [61], [70], [71], [72], and [73], have used profession-based or diagnostic prompts to uncover biases in gender, race, and mental health depictions.

b: EMBEDDING ASSOCIATION TESTS

Embedding association tests (EATs), such as the Word Embedding Association Test (WEAT) and Sentence-Context WEAT (SC-WEAT), measure implicit biases in word

embeddings by computing cosine similarities between social groups and concepts. Caliskan [55] extended these tests to vision and vision-language models, uncovering biases in gender, race, and social representations. This approach has been widely adopted to quantify biases in models like Stable Diffusion and DALL-E, revealing their tendency to amplify stereotypes.

c: HUMAN EVALUATION

Human evaluation involves manual categorization and analysis of generated images by human annotators. This method is particularly effective for assessing subtle biases that automated systems might miss. For instance, Ali et al. [56] used seven independent reviewers to analyze 2,400 images across surgical specialties, comparing them to real-world demographic data. Similarly, Gisselbaek et al. [57] employed human evaluators to categorize AI-generated images of intensivists by sex, age, and race/ethnicity. Other studies, such as [69], [74], [75], and [76], have combined human evaluation with automated metrics to ensure robust bias assessment.

d: AUTOMATED METRICS AND TOOLS

Automated metrics and tools provide scalable and objective measures of bias in TTI models. These include CLIP-based alignment models, Fréchet Inception Distance (FID), and Vision Question Answering (VQA) systems. For example, Luo et al. [58] introduced FAIntbench, a benchmark that uses 18 automated metrics to evaluate bias across 2,654 prompts and 2.1 million images. Similarly, D'Incà et al. [59] developed GradBias, a gradient-based method to quantify how individual words influence bias in generated images. Other tools, such as TIBET [77], OpenBias [78], and [79], leverage large language models (LLMs) and VQA systems to detect and quantify biases dynamically.

e: LATENT SPACE ANALYSIS

Latent space analysis involves examining the internal representations of TTI models to identify biased patterns. Luccioni [60] employed clustering-based evaluation to quantify demographic disparities in Stable Diffusion and DALL·E 2. These methods provide insights into how biases are embedded in the model's latent space and offer opportunities for mitigation.

f: NOVEL FRAMEWORKS FOR BIAS DETECTION

Several studies have proposed novel frameworks for bias detection and quantification. For example, Wang et al. [80] adapted the Implicit Association Test (IAT) to create the Text-to-Image Association Test (T2IAT), measuring implicit associations between concepts and social attributes. Similarly, Seshadri et al. [81] compared Stable Diffusion outputs to training dataset distributions to measure bias amplification. Other frameworks, such as [63] and [82], introduce advanced techniques like cross-attention editing and time-dependent importance reweighting to address intersectional biases.

2) BIAS MEASUREMENT METRICS

Building on the findings from the studies previously reviewed, this subsection examines the metrics used to quantify bias in TTI models. The methodologies employed across these studies reveal a diverse set of approaches for assessing bias, ranging from demographic distribution analysis to embedding-based evaluations, automated classification, statistical scoring, and large-scale benchmarking. Table 4 provides an overview of the various bias measurement metrics identified.

A predominant approach in bias evaluation involves analyzing the demographic distribution of generated images, particularly in terms of gender and racial representations. Several studies [54], [56], [79], [83], [84], [85] measure the proportion of male and female figures or different racial groups in generated outputs and compare them to expected real-world distributions. In cases where equal representation is desired, Mean Absolute Deviation (MAD) is commonly used to quantify deviations from uniform demographic distributions [79], [85].

Other statistical measures include the Neutrality metric [70], which assesses demographic skews in models prompted with bias-neutral descriptions, and the Stereotype Score [86], which quantifies the extent to which certain professions and roles are disproportionately associated with specific demographic groups.

Embedding-based approaches offer an alternative method for measuring bias by analyzing associations in the latent space of multimodal models. Caliskan [55] applies the Word Embedding Association Test (WEAT) and its multimodal variant, SC-WEAT, to identify implicit associations between social concepts and demographic attributes using cosine similarity. Similarly, Wang et al. [80] introduces the Text-to-Image Association Test (T2IAT), adapting the Implicit Association Test (IAT) from psychology to measure biases embedded in TTI model outputs.

Several studies employ automated image classification techniques to quantify bias. BLIP-2 is frequently used to classify gender through visual question answering [79], [87], while FairFace and DeepFace are employed to analyze racial and ethnic representations in generated images [85]. Additionally, Cho et al. [79] incorporates the Face Alignment Network (FAN) and TRUST models to improve skin tone classification.

Object segmentation techniques, such as the Segment Anything Model (SAM), have been utilized to separately evaluate biases in objects and backgrounds within generated images [73], highlighting how models can reinforce stereotypes not only through human depictions but also through scene composition.

Beyond direct classification, some studies introduce statistical scoring mechanisms to assess bias at a systemic level.

TABLE 4. Overview of bias measurement metrics in text-to-image studies, including their scales.

Category	Metrics	Scale	Studies
Distributional Analysis	Gender and Race Distribution	Percentage-based	Masrourisaadat et al. [54], Ali et al. [56], Cho et al. [79], Currie et al. [83], Chauhan et al. [84], Wu et al. [85]
	Mean Absolute Deviation (MAD)	Numerical (Deviation from uniform distribution)	Cho et al. [79], Wu et al. [85]
	Neutrality Metric	Numerical (0 = completely neutral, higher values = more biased)	Sathe et al. [70]
	Stereotype Score	Numerical (higher values = more stereotypical depictions)	Wan and Chang [86]
Embedding-Based BiasMeasures	WEAT, SC-WEAT	Effect size (d)	Caliskan [55]
	Text-to-Image Association Test (T2IAT)	Cosine similarity difference	Wang et al. [80]
Automated Image Classification	BLIP-2 (Gender Classification)	Categorical (Male/Female)	Cho et al. [79], Shin et al. [87]
	FairFace, DeepFace (Face Detection)	Categorical (Gender, Age, Race)	Cho et al. [79], Wu et al. [85]
	Segment Anything Model (SAM)	Object Segmentation (Categorical)	Sureddy et al. [73]
	Mini-InternVL-4B (Demographic Alignment)	Percentage-based	Luo et al. [69]
Statistical Bias Scores	Implicit and Explicit Bias Scores	Numerical (higher values = more biased)	Luo et al. [69]
	Manifestation Factor	Percentage (0-100%)	Luo et al. [69]
	Bias-W (Overall) and Bias-P (Within-image)	Numerical (higher values = stronger bias)	Jiang et al. [67]
Image Diversity & Realism	Fréchet Inception Distance (FID)	Numerical (lower = better image quality)	Ma et al. [88]
	Classification Accuracy Score (CAS)	Percentage-based	Ma et al. [88]
	Vendi Score (Diversity Measure)	Entropy-based diversity measure	Zhang et al. [74]
Bias in Word-to-Image Associations	CLIP-Based Word Attribution	Cosine similarity score	Lin et al. [89]
	GradBias (Gradient-Based)	Numerical (higher values = stronger word-bias correlation)	D'Incà et al. [59]
Intersectional & Socioeconomic Bias	Intersectional Bias (Gender & Race)	Percentage-based comparison	Jiang et al. [67]
	Socioeconomic Representation Bias	Categorical (e.g., wealth/poverty depictions)	Wu et al. [85]
	Caste Bias (Cosine Similarity)	Cosine similarity (higher values = closer to dominant caste representation)	Ghosh [90]
Face-Specific Bias in Generative Models	Face Verification Accuracy (Facenet)	Percentage-based (0-100%)	Rosenberg et al. [91]
	CLIP-Based Face Similarity	Cosine similarity score	Rosenberg et al. [91]
Benchmarks for Large-Scale Bias Evaluation	BIGbench (47,040 prompts)	Multiple metrics (categorical, percentage-based)	Luo et al. [69]
	FAIntBench (2.1 million images)	18 automated bias metrics	Luo et al. [58]
	HEIM (26 Models Evaluated)	CLIPScore, fairness classifiers	Lee et al. [92]

Luo et al. [69] presents implicit and explicit bias scores that measure deviations from real-world demographics, while also introducing a manifestation factor to distinguish between biases arising from a lack of diversity versus those resulting from active stereotyping.

Jiang et al. [67] proposes Bias-W (overall bias) and Bias-P (within-image bias) to quantify disparities in multi-individual images, offering a structured framework for assessing representational imbalances.

Bias assessment also extends to evaluating realism and diversity in generated images. Ma et al. [88] applies Fréchet Inception Distance (FID) and its CLIP-based variant (FIDCLIP) to evaluate the realism and semantic alignment of images. To quantify cultural diversity,

Zhang et al. [74] introduces the Vendi Score, an entropy-based metric that measures how well models capture diverse cultural representations.

In addition, the Classification Accuracy Score (CAS) [88] is used to evaluate the correctness of regional and cultural depictions in generated outputs.

A growing number of studies highlight the importance of measuring intersectional and socioeconomic biases, where multiple demographic attributes influence model outputs. Jiang et al. [67] measures bias at the intersection of gender and race, identifying cases where models amplify disparities across these categories.

Wu et al. [85] extends this analysis to socioeconomic biases, showing how models disproportionately associate racial groups with specific economic conditions, such as linking marginalized groups with poverty while predominantly depicting White individuals in affluent settings.

Ghosh [90] examines caste bias using CLIP-based cosine similarity, revealing disparities in how different caste identities are represented in Stable Diffusion's generated images.

To facilitate systematic bias evaluation across multiple models, some studies introduce large-scale benchmarks. Luo et al. [69] presents BIGbench, a dataset containing 47,040 structured prompts designed to measure biases in gender, race, and occupational depictions.

Luo et al. [58] extends this approach with FAIntBench, a large-scale benchmark applying 18 automated bias metrics to over 2.1 million generated images.

Additionally, Lee et al. [92] introduces the Holistic Evaluation of Text-to-Image Models (HEIM), which integrates CLIP-based fairness metrics with large-scale human evaluations to compare biases across 26 different TTI models.

As summarized in Table 4, the surveyed studies employ a wide range of bias measurement techniques, each addressing different dimensions of bias in generative models. Demographic distribution analysis remains the most widely used approach, but embedding-based evaluations and statistical bias scoring methods provide deeper insights into the underlying biases within TTI models.

While this review identifies a wide range of bias evaluation metrics, it is important to note that the field currently lacks a universally accepted standard for measuring bias in TTI models. This diversity in methodologies—ranging from demographic distribution analyses to embedding association tests and benchmark-based scoring—makes direct comparison across studies challenging. Although we summarize commonly used methods and emerging benchmarks like BIGbench and FAIntbench, a systematic cross-comparison of these metrics is beyond the scope of this review. Nevertheless, our categorization provides a foundation for future work aimed at developing standardized, culturally-aware evaluation frameworks.

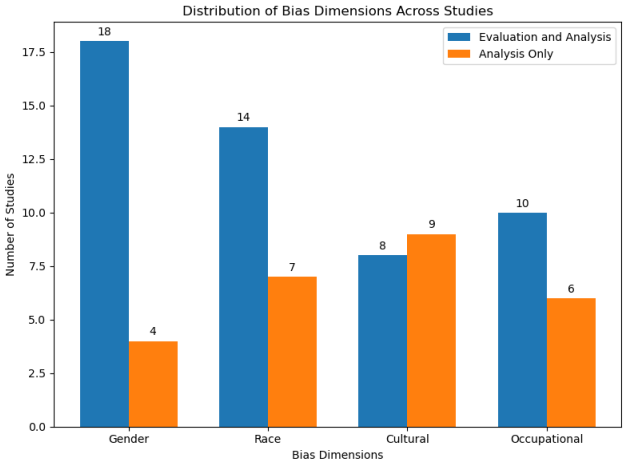

FIGURE 10. Distribution of bias dimensions across studies. The bar chart shows the number of studies focusing on gender, race, cultural, and occupational biases, categorized by study type (Evaluation and Analysis vs. Analysis Only). Gender bias is the most frequently studied dimension, followed by race and occupational biases. Notably, cultural bias—despite being central to this review—is the least addressed in evaluation-driven studies, indicating a methodological gap in how cultural representation is systematically assessed. This underrepresentation suggests an urgent need for developing culturally grounded evaluation tools. (Figure redrawn).

3) MITIGATION METHODS AND EFFECTIVENESS

Recent advances in bias mitigation for TTI models reveal four interconnected paradigms that address different stages of the generation pipeline.

a: LINGUISTIC INTERVENTIONS: REWRITING PROMPTS, REWRITING BIAS

The most accessible approaches target prompt engineering, though their effectiveness varies significantly. Sureddy et al. [73] demonstrated that replacing region-based prompts (e.g., "*bag in Europe*") with adjective-noun constructions ("*European bag*") reduced geographic stereotyping by 52%. While Clemmer et al. [62] automated this process using LLM-based prompt rewriting to achieve **32.8% racial bias reduction**, Bianchi et al. [61] exposed the fragility of manual interventions—explicit counter-stereotypical prompts often failed to override deeply embedded biases. Even advanced frameworks like **FairCritic** [86], which uses GPT-4 Vision feedback loops, risk overcorrection despite reducing occupational stereotype scores from 92 to 26.

b: MODEL-CENTRIC ADJUSTMENTS: TWEAKING ARCHITECTURES

Architectural innovations prioritize surgical precision over scalability. **MIST** [63] reduced intersectional bias by 84% through cross-attention map disentanglement, while plug-and-play **LiVO** [64] encoder slashed gender bias by 42%. Contrastingly, data-centric methods like **Diversity Fine-Tuning (DFT)** [106] improved skin tone fairness by 150% using synthetic datasets—a scalable but computationally intensive approach. The trade-off is stark: **TIW-DSM** [82]

TABLE 5. Summary of bias identification and measurement studies.

#	Authors (Year)	Main Contribution/Methods
		Evaluation and Analysis Studies
1	Wang et al. [80]. (2023)	Introduced the Text-to-Image Association Test (T2IAT), adapting the Implicit Association Test (IAT) to measure biases in TTI models.
2	Cho et al. [79]. (2023)	Automated gender and skin tone detection using BLIP-2 and FAN, introducing Mean Absolute Deviation (MAD) to quantify bias.
3	Caliskan [55]. (2023)	Extended embedding association tests (WEAT, SC-WEAT) to vision-language models, introducing metrics to quantify biases in gender, race, and social representations.
4	Vice et al. [93]. (2023)	Introduced three key metrics (Distribution Bias, Jaccard Hallucination, Generative Miss Rate) to quantify bias in TTI models.
5	Seshadri et al. [81]. (2023)	Compared Stable Diffusion outputs to training dataset distributions to measure bias amplification, introducing controlled prompt variations.
6	Lin et al. [89]. (2023)	Used word-level attribution to identify bias-inducing words in prompts, introducing a pipeline for bias attribution and scoring.
7	Naik and Nushi [94]. (2023)	Proposed a multi-dimensional analysis framework, introducing expanded prompts and automated assessment tools for bias evaluation.
8	Sathe et al. [70]. (2024)	Introduced the Neutrality metric to quantify biases in profession-based prompts, assessing demographic attributes in generated images.
9	Luo et al. [69]. (2024)	Developed BIGbench, a benchmark with 47,040 prompts, introducing implicit and explicit bias scores and a manifestation factor to evaluate bias across models.
10	Sureddy et al. [73]. (2024)	Introduced Decomposed-DIG, a metric to separately evaluate object and background biases in AI-generated images.
11	Luo et al. [58]. (2024)	Developed FAIntbench, a large-scale benchmark with 18 automated metrics to evaluate bias across 2.1 million images.
12	D'Incà et al. [59]. (2024)	Developed GradBias, a gradient-based method to quantify word-level bias influence in TTI models.
13	Lee et al. [92]. (2024)	Conducted holistic evaluation of 26 TTI models using automated metrics (CLIPScore, fairness classifiers) and human evaluations.
14	Zhang et al. [74]. (2024)	Assessed cultural representativeness using qualified human evaluation, patchwork detection, and Vendi scores for diversity analysis.
15	Chinchure et al. [77]. (2024)	Developed TIBET, a framework for dynamic bias detection using GPT-3.5-turbo and Visual Question Answering (VQA).
16	D'Incà et al. [78]. (2024)	Developed OpenBias, a framework to identify biases using LLM analysis, VQA, and entropy-based scoring.
17	Baines et al. [95]. (2024)	Developed a digital storytelling (DST) tool for children, creating a dataset of AI-generated images for bias analysis.
18	Wan and Chang [86]. (2024)	Introduced the Paired Stereotype Test (PST), a novel assessment method to measure gender bias in dual-subject prompts.
19	Kaufenberg-Lashua et al. [96]. (2024)	Proposed standardized prompts and evaluation protocols to assess bias in chemistry-related images.
		Analysis Only Studies
20–22	Mannering [53]. (2023), Sami et al. (2023),Masrourisaadat et al. [54]. (2024)	Analyzed gender bias using paired prompts and human evaluation (e.g., object associations, software engineering roles). Documented male-dominated representations.
23–25	Flathers et al. [72]. (2024), Kuchlous et al. [97]. (2024), Shin et al. (2024)	Investigated model-specific biases in mental health depictions, embedding spaces, and prompt modifiers (e.g., Midjourney, DALL-E).
26–28	Bianchi et al. (2023), Ali et al. [56]. (2024), Ghate et al. [75]. (2024)	Evaluated racial/ethnic disparities in surgical specialties, regional occupations, and language-specific traits.
29–31	Sagar et al. [98]. (2024), Apiola et al. [76]. (2024), Ma et al. [88]. (2023)	Characterized demographic disparities in failure modes, student-reported biases, and food imagery.
32–34	Pal et al. [99]. (2024), Currie et al. [100]. (2024), Fadahunsi et al. [101]. (2025)	Focused on intersectional biases (gender + ethnicity) in face generation, medical roles, and software engineering.
35–37	Chauhan et al. [84]. (2024), Wu et al. [85]. (2024), Ghosh [90]. (2024)	Analyzed race/caste biases in overrepresentation patterns, socioeconomic shifts, and caste stereotypes.
38–40	Rosenberg et al. [91]. (2024), York et al. [102]. (2024), Abrar et al. [103]. (2025)	Documented stereotypical representations in face quality, model comparisons (DALL-E vs. Adobe Firefly), and religious imagery.
41–43	Gisselbaek et al. [57]. (2024), [104]. (2024), [60]. (2024)	Evaluated occupational and cultural biases in medical roles, facial expressions, and cross-profession imbalances.
44–45	[105]. (2024), [68]. (2024)	Analyzed nationality and identity stereotypes using culturally contextual prompts (e.g., visual markers for nationalities).

increased underrepresented subgroup accuracy by 70% but required meticulous timestep-dependent reweighting.

c: POST-HOC CORRECTIONS: FIXING OUTPUTS

Post-generation fixes offer adaptability but inherit model biases. Naseh et al. [65] detected adversarial triggers via latent-space clustering, while Dammu et al. [66] reduced gender disparity by 77% using synthetic corrective images. However, the Gaussian Mixture Model (GMM) approach proposed by Pal et al. [99], despite improving facial diversity, relied on pseudo-labels that perpetuated demographic categorization. The augmentation framework **Chameleon** [107] reduced F1-score disparity for Black individuals from 79% to 27%, treating fairness as an add-on rather than a core design principle.

d: CROSS-MODAL SYNERGIES: BRIDGING MODALITIES

Multimodal alignment strategies address bias as a systemic misalignment. Jiang et al. [67] linked text tokens to biased image regions via **Linguistic-aligned Attention Guidance**, cutting gender bias scores by 76%. Meanwhile, Liu et al. [68]'s fusion of text-image alignment metrics boosted non-Western cultural fidelity by 41%. Hybrid frameworks like **InvDiff** [108] improved fair accuracy to 84.6% but required estimating unlabeled biases—a fundamental limitation for real-world deployment.

e: CRITICAL SYNTHESIS

Current methods span a spectrum from *reactive* (post-hoc corrections) to *proactive* (architectural redesigns). While linguistic interventions like prompt engineering lack robustness [87], hybrid approaches combining **MIST**'s attention-level adjustments with **DFT**'s data diversity show promise. However, as Abrar et al. [103] cautioned, no method fully eradicates bias—especially against entrenched stereotypes. Ultimately, effective mitigation requires continuous sociotechnical dialogue between model outputs and human values.

4) CHALLENGES AND LIMITATIONS

Although considerable progress has been made in both measuring and mitigating cultural bias in TTI models, several challenges persist. These challenges stem from methodological, technical, and practical limitations that hinder the deployment of truly fair and robust systems in real-world applications. Fig.11 showcases the number of selected studies per mitigation method and Table.6 summarizes the selected reviewed mitigation focused studies.

- **Inconsistency in Bias Elicitation and Measurement:** Methods based on prompt engineering—such as using triggering terms or counter-stereotypical prompts [61], [89], [103]—often produce inconsistent outcomes. Automated metrics (e.g., DiffusionITM [110], MAD [85], and gradient-based attribution [59]) may fail to capture nuanced cultural contexts, leading to varying assessments of bias.
- **Limitations of Evaluation Frameworks:** Large-scale benchmarks like BIGbench [69] and FAIntbench [58] provide valuable insights, yet they are often constrained by synthetic prompt designs or limited demographic representations. Moreover, human evaluations—while rich in cultural insight—are subjective and difficult to scale reliably [76], [111].
- **Trade-offs in Mitigation Strategies:** Fine-tuning and latent space interventions (e.g., LoRA fine-tuning [98] and gradient-based debiasing [112]) have demonstrated significant bias reduction. However, these methods can compromise image quality or disrupt identity consistency. In some cases, post-hoc approaches like LLM-based prompt rewriting [62] may mitigate surface-level biases without addressing deeper systemic issues.
- **Scalability and Computational Cost:** Advanced techniques, such as deep reinforcement learning for bias characterization [98] and multi-modal evaluation frameworks [92], demand high computational resources and complex processing pipelines. These resource-intensive methods pose challenges for integration into real-time or large-scale applications.
- **Adversarial Bias and Dynamic Cultural Norms:** Models are vulnerable to adversarial attacks that inject subtle biases [65]. Furthermore, cultural bias is inherently dynamic and context-dependent, making it difficult to develop universal fairness metrics or mitigation strategies that remain valid over time.
- **Practical Adoption Considerations:** Many of the proposed methods work well in controlled experimental settings but face hurdles in real-world deployment.

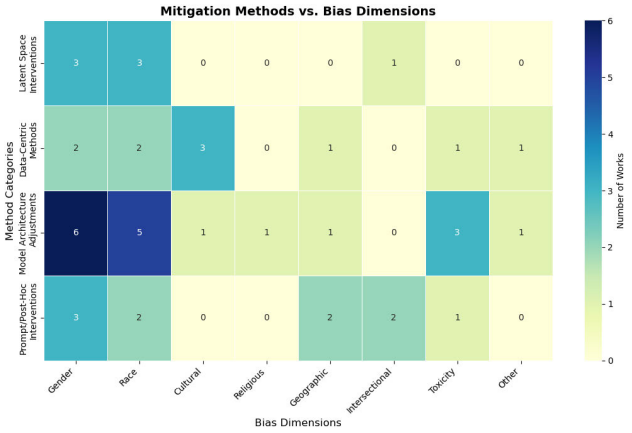

FIGURE 11. Heat map showing the number of works in each mitigation method category addressing specific bias dimensions. Color intensity corresponds to the number of studies. Model architecture adjustments and data-centric methods appear to dominate mitigation efforts for gender and racial bias, while cultural and intersectional biases are comparatively under-addressed across all method categories. This pattern highlights a research imbalance, where technically tractable dimensions (e.g., gender, race) are prioritized, potentially at the expense of more complex socio-cultural factors. (Figure redrawn).

TABLE 6. Summary of bias mitigation methods.

#	Method	Study	Mechanism	Effectiveness
1	Manual counter-prompts	Bianchi et al. [61]. (2023)	Explicitly instructing models to counteract stereotypes (e.g., "a white poor person")	Failed to override racial associations; bias persisted
2	Diversity Fine-Tuning	Esposito et al. [106]. (2023)	Fine-tuning on 89k synthetic images with balanced demographics (ethnicity, gender, age)	150% fairness improvement for skin tone; 97.7% for gender
3	Synthetic correction	Dammu et al. [66]. (2023)	Generating corrective images via DALL-E/Stable Diffusion to retrain classifiers	77% ↓ gender disparity; 50% ↑ accuracy for PoC female doctors
4	Prompt restructuring	Sureddy et al. [73]. (2024)	Replacing region-based prompts ("X in Y") with adjective-noun phrasing ("Y-style X")	52% ↑ background diversity for Africa; 20% avg. improvement
5	LLM fairness rewriting	Clemmer et al. [62]. (2024)	Rewriting prompts via fine-tuned LLM to balance gender/ethnicity terms	32.8% ↓ racial bias; 64.2% ↑ skin tone diversity
6	FairCritic feedback	Wan and Chang [86]. (2024)	GPT-4 Vision critiques generations, iteratively refining prompts	Stereotype score ↓ 92 → 26 (occupations); power bias ↓ 18.98 → -11.12
7	Prompt sequencing	Shin et al. [87]. (2024)	Testing modifier order sensitivity (e.g., "female CEO" vs. "CEO, female")	Inconsistent bias reduction; DALL-E diversity ↑ with modifier-first prompts
8	MIST	Yesiltepe et al. [63]. (2024)	Editing cross-attention maps to disentangle gender/race/age attributes	Bias score ↓ 0.86 → 0.14 (WinoBias); preserved image quality
9	LiVO	Wang and Specia [109]. (2024)	Plug-in encoder injecting ethical principles via diffusion-specific loss	Gender bias ↓ 56.27 → 33.69; toxicity ↓ 66%
10	TIW-DSM	Kim et al. [82]. (2024)	Time-dependent reweighting to approximate unbiased data distribution	Underrepresented subgroups ↑ 12.4% → 21.1% (CelebA)
11	MoESD	Wang and Specia [109]. (2024)	Mixture-of-Experts with bias-expert adapters for targeted debiasing	Fairness score ↓ 0.209 → 0.127 (skin tone); 5.6% params tuned
12	Linguistic alignment	Jiang et al. [67]. (2024)	Aligning attention maps to debias multi-individual scenes	Gender Bias-W ↓ 0.338 → 0.080; racial Bias-P ↓ 0.371 → 0.367
13	Chameleon	Erfanian et al. [107]. (2024)	Augmenting underrepresented groups via guided synthetic generation	F1-score disparity ↓ 79% → 27% (Black individuals)
14	LoRA fine-tuning	Sagar et al. [98]. (2024)	Low-Rank Adaptation on balanced DALL-E3 images	Male-to-female ratio ↓ 1.65 → 1.16; ambiguity ↓ 43%
15	GMM disentanglement	Pal et al. [99]. (2024)	Clustering latent codes for age/gender/race via Gaussian mixtures	Smile classification bias ↓ 65%; accuracy ↑ 93.08% → 94.84%
16	InvDiff	Hou et al. [108]. (2024)	Two-stage invariant guidance to filter spurious correlations	Fair accuracy ↑ 67.2% → 84.6% (Waterbirds)
17	CLIP refinement	Kuchlous et al. [97]. (2024)	Correcting CLIP embeddings for fairer alignment evaluation	Improved fairness in generation & evaluation metrics
18	Backdoor detection	Naseh et al. [65]. (2024)	Clustering latent embeddings to detect adversarial triggers	Systematic identification of backdoor-induced biases
19	Cultural filtering	Liu et al. [68]. (2024)	Curating training data via text-image-object alignment metrics	Cultural fidelity ↑ 41% (C3 benchmark)
20	Positive augmentation	Abrar et al. [103]. (2025)	Adding neutral descriptors (e.g., "peaceful leader") to prompts	Partial reduction in religious stereotypes; DALL-E3 > SD3

Integration with existing systems, the opacity of proprietary models, and the need for real-time processing are significant concerns. Additionally, regulatory and ethical considerations, along with varying definitions of fairness across cultures, further complicate practical adoption [70], [86].

While various bias mitigation strategies have shown promising results in experimental settings, their real-world feasibility and long-term effectiveness remain complex and context-dependent. Techniques such as prompt engineering or cross-modal alignment are relatively lightweight and easier to integrate in end-user systems. However, architectural modifications or synthetic data augmentation often demand significant computational resources and technical expertise, limiting their adoption in production environments.

Moreover, strategies that rely on synthetic balancing or fairness optimization may unintentionally introduce new biases or overcorrect existing ones, especially in culturally nuanced or intersectional scenarios. For example, applying a Western-centric fairness definition may erase meaningful cultural expressions in non-Western contexts. To ensure long-term impact, these methods must be continuously monitored and refined through feedback loops involving diverse stakeholders, including domain experts, cultural practitioners, and affected communities.

From an implementation standpoint, periodic audits, transparent reporting, and modular fairness layers (e.g., plug-and-play encoders or decoders) can enhance scalability and trust. It is also crucial to adapt mitigation strategies to application-specific needs—e.g., educational vs. commercial imagery—and regulatory constraints across regions. Future

work should explore these socio-technical dimensions more deeply, especially under evolving geopolitical and cultural contexts. In summary, while multi-faceted approaches to bias measurement and mitigation have led to promising improvements, the challenges of inconsistency, scalability, and evolving cultural norms remain substantial. Addressing these issues requires not only technical innovations but also interdisciplinary collaboration and ongoing evaluation to ensure that debiased TTI models can be effectively and ethically deployed in diverse real-world contexts.

V. THEMATIC FINDINGS

In this section, we detail the findings from the selected studies in a thematic fashion, where each theme corresponds to one of the main bias dimensions in our systematic review.

1) GENDER BIAS

Throughout this review, it is found that gender bias is one of the most prevalent and well-documented biases in text-to-image (TTI) models. Across multiple studies, these models consistently reinforce traditional gender norms, overrepresent men in high-status professions, and exaggerate occupational stereotypes, even when prompts are gender-neutral.

A dominant theme in the literature is occupational gender bias. Studies show that models overwhelmingly generate male figures for high-status professions and female figures for caregiving or service-oriented roles. When prompts such as ''a doctor at work'' or ''a scientist in a laboratory'' are used without specifying gender, male figures dominate the outputs. Sami et al. [71] found that only 2% of 2,280 images generated for ''As a software engineer'' featured women. Similarly, Fadahunsi et al. [101] reported that all Stable Diffusion models disproportionately depicted male software engineers, with SD 2 exhibiting the strongest bias, followed by SD 3 and SD XL.

Apiola et al. [76], Chinchure et al. [77], Wang et al. [80], Lee et al. [92], Sagar et al. [98], and Luccioni et al. [60] further confirm that models like Stable Diffusion, Midjourney, and DALL-E overrepresent men in leadership, STEM, and labor-intensive roles while depicting women in caregiving and secondary positions. Cho et al. [79], Shin et al. [87], Vice et al. [93] observed that nurses and receptionists are overwhelmingly female, while engineers and mechanics are almost exclusively male. Currie et al. [83], [100], [113] found that DALL-E 3 depicted pharmacists as 69.7% male and cardiologists as 86% male, despite real-world statistics being closer to gender parity. Additionally, Wan and Chang [86] found that 74% of generated images reinforced gender-stereotypical occupations, with CEOs and managers appearing as men and assistants and interns as women. Bianchi et al. [61] found that 99% of AI-generated software developer images depicted white men, despite the actual U.S. workforce being significantly more diverse. Similarly, Bianchi et al. [61] reported that housekeepers were almost always non-white women, while software developers were nearly all white men. In addition, Seshadri et al. [81] showed that while LAION's dataset contains 25% female engineers, Stable Diffusion's generated images reduced that figure to only 10%, amplifying real-world disparities. Luo et al. [58] found that SDXL, PixArt, and Stable Cascade produced balanced representations in neutral settings but still reinforced occupational stereotypes.

Linguistic factors further reinforce gender bias in AI-generated images. D'Incà et al. [59] found that replacing gender-specific words with neutral terms like ''person'' led to more balanced outputs. while Wang et al. [80] and Lin et al. [89] reported that words like ''leader'' strongly favored male depictions, while ''caring'' increased female representation. Ghate et al. [75] additionally highlighted how language also influences gender bias across translations. In Hindi, models generated disproportionately more male figures for traits such as ''business'' and ''hardworking'' compared to English. Additionally, English prompts produced professional office environments, while Hindi prompts led to images of men engaged in physical labor. More over, [95] revealed that Stable Diffusion 2.0 reinforced gender stereotypes in children's storytelling, depicting men in leadership roles while women were placed in caregiving or secondary roles.

Gender bias in TTI models intersects with racial disparities, compounding representational imbalances. Reference [96] found that DreamStudio (Stable Diffusion) generated chemists at a 3:1 male-to-female ratio and overwhelmingly depicted white individuals. Reference [54] found that Stable Diffusion disproportionately generates white males in leadership roles, even when given neutral prompts.

Biases also extend to emotional expression and storytelling. Reference [104] found that DALL-E 2 not only underrepresents women in male-dominated fields but also exaggerates gendered presentational biases. Women were more likely to be smiling and have their heads pitched downward, particularly in traditionally female-dominated roles.

Overall, TTI models do not merely reflect societal gender norms but actively amplify them. Bias is evident in occupational portrayals, language-driven disparities, racial representation, and presentational styles. Despite attempts at fairness constraints, deeply ingrained biases persist, highlighting the need for improved dataset curation, bias-aware model training, and interventions that challenge entrenched stereotypes in AI-generated imagery.

2) RACIAL/ETHNIC BIAS

Text-to-image (TTI) models frequently exhibit racial and ethnic biases, often reinforcing existing stereotypes or underrepresenting certain demographic groups.

Luo et al. [69] found that these models struggle with racial pairings in relational prompts, often failing to generate racially diverse compositions. For instance, they were less likely to generate an East Asian husband with a White

TABLE 7. Intersectionality of biases dimensions in TTI models.

	Gender Bias	Racial/Ethnic Bias	Class Bias	Age Bias	Cultural Bias	Geographic Bias
Gender Bias	-	Black women nearly absent in high-status jobs; White men dominate. [Luccioni et al. [60]; Lin et al. [89]; Bianchi et al. [61]]	Women overrepresented in low-income caregiving roles; men in high-income jobs. [Chauhan et al. [84]; Apiola et al. [76]; Luo et al. [69]]	Older women underrepresented in leadership; older men more likely to be in labor-intensive roles. [Luo et al. [58]; Gisselbaek et al. [57]]	Women depicted in caregiving and service roles per cultural norms; gender roles vary across societies. [Shin et al. [87]; Ghate et al. [75]]	Gender distribution varies by region (e.g., pharmacists: female in EU, male in Asia/Africa). Chinchure et al. [77]
Racial/Ethnic Bias	Black women nearly absent in high-status jobs; White men dominate. [Luccioni et al. [60]; Lin et al. [89]; Bianchi et al. [61]]	-	Black and Hispanic individuals in low-income roles; White in high-status jobs. [Chauhan et al. [84]; Luo et al. [58]]	Older marginalized groups shown in labor-intensive jobs. Gisselbaek et al. [57]	Racial groups tied to specific cultural settings (e.g., Asians in traditional depictions); non-Western cultures underrepresented. [Shin et al. [87]; Zhang et al. [74]]	Non-Western racial groups underrepresented/misrepresented; Western beauty standards dominate. [Vice et al. [93]; Liu et al. [68]]
Class Bias	Women overrepresented in low-income caregiving roles; men in high-income jobs. [Chauhan et al. [84]; Apiola et al. [76]; Luo et al. [69]]	Black and Hispanic individuals in low-income roles; White in high-status jobs. [Chauhan et al. [84]; Luo et al. [69]]	-	Older individuals associated with poverty and service jobs. [Luo et al. [69]; Gisselbaek et al. [57]	African settings shown as poor; Western as wealthy. [Wu et al. [85]; Sureddy et al. [73]]	Low-income depictions skew toward non-Western locations. Vice et al. [93]
Age Bias	Older women underrepresented in leadership; older men more likely to be in labor-intensive roles. [Luo et al. [69]; Gisselbaek et al. [57]]	Older marginalized groups shown in labor-intensive jobs. Gisselbaek et al. [57]	Older individuals associated with poverty and service jobs. Luo et al. [58]	-	Older individuals depicted with wisdom/fragility stereotypes; Western leadership roles favor youth. [Luo et al. [69]; Chinchure et al. [77]]	Older people depicted in lower-income roles in non-Western settings. Wu et al. [85]
Geographic Bias	Gender distribution varies by region (e.g., pharmacists: female in EU, male in Asia/Africa). Chinchure et al. [77]	Non-Western racial groups underrepresented/misrepresented; Western beauty standards dominate. [Vice et al. [93]; Liu et al. [68]]	Low-income depictions skew toward non-Western locations. Vice et al. [93]	Older people depicted in lower-income roles in non-Western settings. Wu et al. [85]	Western-centric depictions dominate; non-Western cultures misrepresented. Zhang et al. [114]	-
Cultural Bias	Women depicted in caregiving/service roles per cultural norms; gender roles vary across societies. [Shin et al. [87]; Ghate et al. [75]]	Racial groups linked to cultural settings (e.g., Asians in traditional depictions); non-Western cultures underrepresented. [Shin et al. [87]; Zhang et al. [74]]	African settings shown as poor; Western settings depicted as wealthy. [Wu et al. [85]; Sureddy et al. [73]]	Older individuals depicted with wisdom/fragility stereotypes; Western leadership roles favor youth. [Luo et al. [69]; Chinchure et al. [77]]	-	Western-centric depictions dominate; non-Western traditions misrepresented. [Zhang et al. [74]; Vice et al. [93]]

wife, while the inverse pairing was more reliably produced, reflecting asymmetric stereotype reinforcement.

Kuchlous et al. [97] noted that racial bias in TTI models is among the most difficult to mitigate, often persisting

despite intervention strategies. Additionally, [69] found that distilled models (e.g., SDXL-Lightning, LCM-SDXL) inherit and even exacerbate racial biases from their base models, suggesting that biases are intensified during the distillation process rather than being neutralized.

Studies such as Shin et al. [87] highlight the persistence of racial stereotypes, with models defaulting to culturally ingrained assumptions. For example, monks were overwhelmingly depicted as Asian males, and racial diversity was often difficult to adjust even with explicit prompting. Additionally, models exhibited a Northern Hemisphere bias in seasonal depictions, such as associating ''January'' exclusively with winter. While some models, such as DALL·E, showed more balanced outputs, this raised questions regarding whether AI should reflect demographic realities or promote representational inclusivity.

Skin tone bias remains a critical issue. Cho et al. [79] found that TTI models underrepresent darker skin tones, with most images clustering around mid-range values on the Monk Skin Tone (MST) scale, avoiding both very dark and very light tones. Similarly, Ali et al. [56] showed that surgeons were overwhelmingly depicted as White and male, despite real-world diversity, and that geographic-based prompting increased non-White representation but failed to improve gender diversity.

Bianchi et al. [61] found strong racial and geographic biases in depictions of poverty, crime, and economic status. Neutral prompts such as ''a poor person'' disproportionately generated Black individuals, while ''a terrorist'' overwhelmingly depicted brown-skinned men with beards, reinforcing racial profiling. Likewise, when generating images of African men with cars, models often depicted old or damaged vehicles, while American men were associated with luxury cars. Even when countering stereotypes with explicit modifiers (e.g., ''a wealthy African man''), models continued to produce stereotypical imagery, indicating deep-seated biases in training data.

Professional and occupational biases also reflect racial disparities. Reference [58] found that SDXL, PixArt, and Stable Cascade favored White individuals in high-status roles while disproportionately associating darker-skinned individuals with low-income or negative social contexts. Similarly, [76] found that doctors and professors were mostly depicted as White, while criminals and terrorists were disproportionately shown as dark-skinned men.

Reference [101] identified significant underrepresentation of Black and Arab individuals across all models. Even with diverse prompting strategies, SD 2 and SD XL consistently overrepresented White figures, while SD 3 slightly favored Asian figures but still neglected Black and Arab populations. This suggests that racial disparities in AI-generated imagery persist across multiple iterations and architectures.

These biases are not static but shift over time. Wu et al. [85] found that negative stereotypes (e.g., depicting criminals and poor individuals as Black or Latino) decreased in later Stable Diffusion models but shifted toward Asian individuals. For example, in SD-XL, 70.5% of images generated for ''poor person'' depicted Asians, compared to only 25.5% in SD-1.5. Meanwhile, White individuals remained largely unaffected by negative stereotypes across all model versions.

The quality of AI-generated images also varies by racial representation. Rosenberg et al. [91] found that White individuals consistently received higher-quality images, while Black, East Asian, and Indian individuals had lower-rated outputs. Face verification accuracy was lower for synthetic faces of non-White individuals, highlighting disparities in representation and data quality.

3) SOCIAL BIASES

TTI models often encode biases related to social attributes such as occupation, class, and age. Occupational stereotypes are particularly pervasive, with AI-generated images frequently overrepresenting certain demographic groups in specific professions.

Reference [58] found that SDXL, PixArt, and Stable Cascade reinforce traditional occupational stereotypes. High-status professions, such as CEOs, scientists, and software engineers, are predominantly depicted as male and White, while service and caregiving roles, such as nurses and teachers, are disproportionately assigned to women and people of color.

Reference [76] highlighted age-related biases, with business leaders consistently depicted as older White men, while younger individuals were more frequently shown in creative and digital professions. Similarly, [57] found that AI-generated images overrepresented young intensivists (<40 years old) while significantly underrepresenting older professionals (>60 years old).

Class-based biases also emerge in AI-generated imagery. Reference [73] found that images of low-income individuals in non-Western regions often default to rural, underdeveloped settings, reinforcing stereotypes about economic status and geographic development. Likewise, [77] found that models disproportionately associate wealth with White individuals, while depictions of non-Western individuals in professional settings were limited.

4) CULTURAL ASSOCIATIONS AND INTERSECTIONALITY

Cultural associations and cultural representation in TTI models is often skewed, favoring dominant Western narratives while misrepresenting or marginalizing other cultures. Reference [87] reported that DALL.E- and Stable Diffusion frequently reinforced dominant cultural associations, such as linking ''Lunar New Year'' exclusively to China or monks to Asian depictions. Attempts at diversification, particularly by Firefly, sometimes conflicted with demographic authenticity, raising ethical concerns about bias mitigation.

Reference [88] analyzed culinary culture in AI-generated images and found that Stable Diffusion, mini DALL.E, and DALL.E small exhibited significant bias, with Asian cuisine receiving less accurate representation compared to European, North American, and Latin American cuisines. This suggests

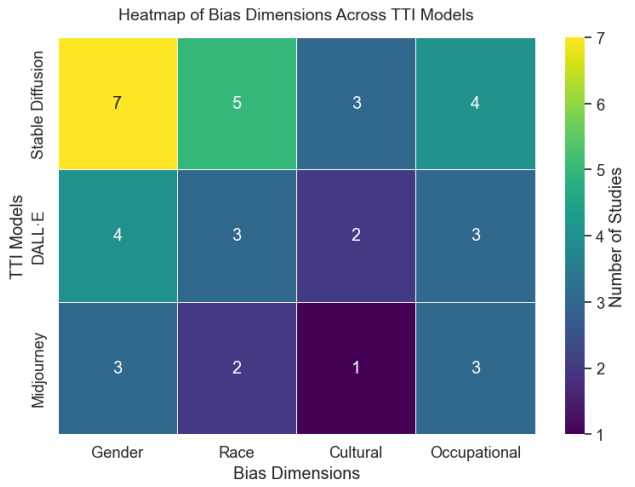

FIGURE 12. Heatmap of bias dimensions across TTI models. The heatmap visualizes the frequency of studies analyzing four primary bias dimensions—gender, race, cultural, and occupational—in three prominent TTI models: Stable Diffusion, DALL-E, and Midjourney. Each cell indicates the number of studies addressing a specific bias dimension for a given model, with brighter shades representing higher frequencies. Stable Diffusion is the most scrutinized model, particularly in relation to gender and race, suggesting its centrality in fairness research. In contrast, DALL-E and Midjourney have received comparatively less attention across all dimensions. This disparity highlights a need for broader evaluation of cultural and occupational biases across diverse model architectures. (Figure redrawn).

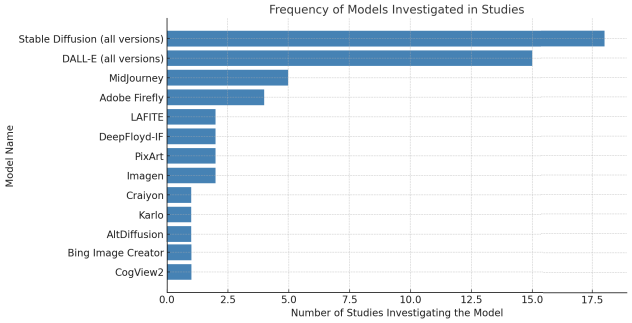

FIGURE 13. Bar chart illustrating the frequency of models' Investigation in the selected studies. The figure reveals that Stable Diffusion (all versions) and DALL·E (all versions) are the most frequently examined models, significantly outpacing others such as Midjourney and Adobe Firefly. Less commonly analyzed models include Craiyon, Karlo, and CogView2, indicating a potential gap in comparative analysis across less popular models.

TABLE 8. Bias amplification trends observed across Stable Diffusion versions. Studies indicate that gender bias increased significantly in later versions [85], [101], while racial bias shifted toward different demographic groups rather than being mitigated [69], [85]. Age bias, initially less prominent, became more evident in SD-XL [58], [77], and cultural bias exhibited the strongest increase across versions [68], [74].

Bias Type	SD 1.5 → SD 2.1	SD 2.1 → SD XL
Gender Bias	Increased [84, 85]	Stronger increase [85, 101]
Racial Bias	Shifted [69, 85]	Shifted toward Asian individuals [85]
Age Bias	Emerging [58, 77]	Significant increase [58]
Cultural Bias	Increased [73, 85]	Highest increase [68, 74]

a training data imbalance favoring Western cultural elements over non-Western ones.

Stereotypical geographic portrayals are also common. Reference [73] reported that generated images of Africa predominantly featured rural landscapes with dirt roads, whereas European settings often showcased historical stone architecture. Similarly, [85] found that images of African individuals frequently depicted poverty-related stereotypes, such as tattered clothing and deteriorating infrastructure. Even when countering stereotypes with explicit prompts, biases were not fully eliminated. Intersectionality compounds these biases. Reference [60] found that gender and racial disparities intersect in occupational depictions, with Black women almost entirely absent from executive positions, while White men dominated portrayals of CEOs and engineers. Reference [77] showed that regional biases influenced gender distributions, such as depicting pharmacists in Europe as predominantly female, while in Asia and Africa they were mostly male. Table.7 reports intersectionality between different bias dimensions that have been reported in the selected studies.

These biases influence cultural perception, reinforcing stereotypes about social roles and economic status across different regions. The entanglement of gender, race, and cultural associations suggests that TTI models do not merely reflect biases in their training data but actively reinforce existing societal hierarchies.

Trends in TTI Models: The TTI models that have been investigated in the selected literature vary in their count, performance, and evolution; here we provide a number of insights into them. Fig.13 reports how many times each model has been investigated in a selected study. Table.8 shows the performance of the Stable Diffusion model over 3 of its versions across different bias dimensions as reported in several selected studies.

A. IDENTIFIED GAPS

Although research on bias identification and mitigation in TTI models has expanded considerably, several critical gaps remain:

- **Diversity in Training Data and Benchmarks:** Existing datasets and benchmarks—such as those discussed in [68] and [74]—are predominantly Western-centric, limiting the representation of global cultural nuances and diverse demographic groups.
- **Standardized Evaluation Metrics:** While various quantitative and qualitative metrics have been proposed (e.g., the DiffusionITM score [110] and Mean Absolute Deviation [85]), there is no unified framework that comprehensively captures bias across dimensions such as gender, race, age, and class.
- **Trade-offs in Mitigation Strategies:** Methods like fine-tuning [106] and latent space interventions [112] show promise in reducing bias; however, they often incur trade-offs such as diminished image quality or disrupted

identity consistency. Moreover, many strategies target single bias dimensions rather than the multifaceted nature of real-world biases.

- **Underexplored Intersectionality:** Although initial insights into intersectional biases are provided by studies such as [60], [77], and [89], the complex interplay between gender, race, geography, and age is still insufficiently examined. More work is needed to understand how these factors interact and compound one another.
- **Dynamic Bias Shifts:** As models are updated, some biases are not reduced but instead shift across dimensions [85]. Tracking and mitigating these evolving bias patterns remain an open challenge.
- **Scalability and Practical Integration:** Many of the current evaluation and mitigation techniques demand substantial computational resources or manual oversight, posing difficulties for their application in real-time or large-scale, real-world scenarios.

Despite ongoing advances, there is currently no definitive solution to mitigating bias in TTI models. Existing methods—such as dataset curation, fine-tuning, and prompt engineering—offer partial improvements but fail to fully address the structural and cultural dimensions of bias. This highlights a critical gap: the need for holistic, multilevel strategies that integrate technical, cultural, and ethical perspectives to achieve more sustainable and equitable outcomes.

A notable gap in current bias evaluation practices is the Western-centric orientation of most existing benchmarks. Tools such as BIGbench and FAIntbench, while useful, are largely constructed around Western cultural norms, language, and societal structures. This focus limits their effectiveness in detecting biases that affect non-Western or underrepresented cultural contexts. As a result, models may appear unbiased when assessed through these benchmarks but still propagate cultural stereotypes or omissions when generating content related to other regions. This points to a critical need for the development of culturally inclusive benchmarks and evaluation protocols that reflect the diversity of global perspectives and experiences. Without this expansion, fairness assessments in TTI models will remain incomplete and potentially misleading.

Addressing these gaps is essential for developing robust, culturally aware, and practically viable TTI systems that accurately reflect the diverse tapestry of human society.

VI. DISCUSSION

A. INTERPRETATION OF FINDINGS

The body of research on cultural bias in TTI models reveals that these systems consistently reproduce and often amplify societal stereotypes. Studies have documented pervasive biases across multiple dimensions—including gender, race, age, and class—that not only mirror existing cultural hierarchies but also reinforce them. For example, numerous investigations have shown that models tend to overrepresent dominant cultural narratives (e.g., White, male, Western aesthetics) while marginalizing non-Western and minority groups [61], [68], [100].

A promising avenue to enhance and standardize bias evaluation is the development of Cultural Relevance Index (CRI) [115]. By integrating multidimensional metrics—such as those found in benchmarks like BIGbench [69] and FAIntbench [58]—CRI can offer a unified framework for quantifying bias across diverse cultural, demographic, and geographic variables. Standardization through CRI holds the potential to not only compare models more effectively but also to guide developers in systematically addressing cultural misrepresentations.

Although most studies analyze bias as a static issue, cultural prejudice is in fact dynamic and influenced by evolving societal norms, media narratives, and political climates. As socio-cultural contexts shift, so too do the ways in which prejudice manifests in TTI outputs. Current models are often trained on static datasets that fail to reflect these ongoing changes, leading to outdated or mismatched visual representations. Therefore, we emphasize the need for longitudinal studies and adaptive model design that can capture and respond to the temporal evolution of cultural norms. Incorporating continuous learning strategies and regularly updating training data with culturally relevant content can help TTI systems remain aligned with current values and reduce the entrenchment of obsolete stereotypes.

B. BROADER IMPLICATIONS

The ethical and societal implications of biased TTI models are profound. When AI systems reinforce harmful stereotypes—such as associating leadership exclusively with White men or reducing non-Western cultural traditions to simplistic or impoverished images—they risk perpetuating inequality and further marginalizing already underrepresented groups. The erasure or misrepresentation of diverse cultural identities in AI-generated imagery can distort public perceptions and influence decision-making processes in areas such as media, education, and employment.

In response to these challenges, there is a growing need for the implementation of policy and ethical frameworks that can govern the responsible development and deployment of TTI models. Ethical AI design should be rooted in principles of transparency, inclusivity, accountability, and cultural sensitivity. This includes adopting auditing tools that evaluate representation quality across cultural contexts, enforcing transparency in data sourcing and labeling, and ensuring explainability of generative outputs.

On the policy level, regulators and industry bodies can play a pivotal role by defining culturally-aware standards and enforcing compliance through certification mechanisms, similar to existing practices in domains like data privacy or medical AI. Collaboration between policymakers, ethicists, and technologists is essential for establishing ethical

guidelines that evolve with cultural contexts. Sector-specific applications of TTI (e.g., in advertising, education, or media) may also benefit from dedicated codes of conduct to prevent cultural misrepresentation.

In response to these challenges, there is significant potential for policy makers and industry stakeholders to develop guidelines that promote fairness and accountability in AI. Establishing clear standards for cultural representation and bias mitigation in TTI systems can drive the adoption of ethical practices across the industry. These standards could be incorporated into regulatory frameworks or best-practice guidelines, ensuring that developers prioritize culturally sensitive data collection, transparent evaluation methods like CRI, and continuous monitoring of model outputs.

Despite the proliferation of evaluation methods, the absence of standardized bias measurement practices limits comparability across studies. We recognize the value of conducting a formal comparison of these metrics; however, such an empirical undertaking—requiring consistent experimental setups and reimplementation of diverse methodologies—is beyond the scope of this review. Instead, we highlight the urgent need for community consensus around culturally inclusive benchmarks and recommend future research toward harmonizing evaluation protocols.

In summary, while current research has made valuable strides in identifying and quantifying cultural biases in TTI models, there remains a critical need for standardized evaluation processes and robust mitigation strategies. These efforts must be complemented by policy and ethical frameworks that govern TTI deployment at scale. The broader societal and ethical stakes demand coordinated efforts from researchers, developers, and policy makers to ensure that AI technologies contribute to a more inclusive and equitable digital landscape.

VII. CONCLUSION

This systematic review synthesized a broad range of studies investigating bias identification and mitigation in TTI models. The findings indicate that these models consistently reproduce and, in many cases, amplify cultural biases across multiple dimensions—including gender, race/ethnicity, age, and class—and that such biases often intersect, compounding the challenges of representation.

Despite promising advances in mitigation strategies, such as fine-tuning, latent space modifications, and prompt-based interventions, no single approach has yet achieved comprehensive fairness without trade-offs in image quality or representational accuracy. The potential development of standardized evaluation frameworks could possibly offer a more unified and systematic means of measuring and addressing these biases.

The ethical and societal implications of biased TTI models are significant. By reinforcing harmful stereotypes and misrepresenting diverse cultural narratives, these models risk perpetuating inequality and narrowing our collective understanding of global cultures. Future research must address the identified gaps—ranging from the need for culturally diverse training data to more nuanced methods for capturing intersectional bias—while also exploring dynamic mitigation strategies that adapt to evolving cultural contexts.

In summary, while considerable progress has been made in understanding and mitigating biases in TTI models, substantial challenges remain. Addressing these issues is critical to ensuring that AI-driven creative tools contribute to a more equitable and inclusive digital landscape.

ACKNOWLEDGMENT

Open Access funding provided by the Qatar National Library.

REFERENCES

[1] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.

[2] H.-K. Ko, G. Park, H. Jeon, J. Jo, J. Kim, and J. Seo, "Large-scale text-to-image generation models for visual artists' creative works," in *Proc. 28th Int. Conf. Intell. User Interface*, Mar. 2023, pp. 919–933.

[3] M. Mandal, D. Ghadiyaram, D. Gurari, and A. C. Bovik, "Helping visually impaired people take better quality pictures," *IEEE Trans. Image Process.*, vol. 32, pp. 3873–3884, 2023.

[4] L. Höllein, A. Cao, A. Owens, J. Johnson, and M. Nießner, "Text2Room: Extracting textured 3D meshes from 2D text-to-image models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7875–7886.

[5] H. Vartiainen and M. Tedre, "Using artificial intelligence in craft education: Crafting with text-to-image generative models," *Digit. Creativity*, vol. 34, no. 1, pp. 1–21, Jan. 2023.

[6] V. Vimpari, A. Kultima, P. Hämäläinen, and C. Guckelsberger, "'An adapt-or-die type of situation': Perception, adoption, and use of text-to-image-generation AI by game industry professionals," *Proc. ACM Hum.-Comput. Interact.*, vol. 7, pp. 131–164, Jan. 2023.

[7] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 36479–36494.

[8] D. Yu, X. Li, C. Zhang, T. Liu, J. Han, J. Liu, and E. Ding, "Towards accurate scene text recognition with semantic reasoning networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12110–12119.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.

[11] R. Daneshjou, M. P. Smith, M. D. Sun, V. Rotemberg, and J. Zou, "Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review," *JAMA Dermatol.*, vol. 157, no. 11, pp. 1362–1369, 2021.

[12] T. P. Pagano, R. B. Loureiro, F. V. N. Lisboa, R. M. Peixoto, G. A. S. Guimarães, G. O. R. Cruz, M. M. Araujo, L. L. Santos, M. A. S. Cruz, E. L. S. Oliveira, I. Winkler, and E. G. S. Nascimento, "Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," *Big Data Cognit. Comput.*, vol. 7, no. 1, p. 15, Jan. 2023.

[13] C. Bird, E. Ungless, and A. Kasirzadeh, "Typology of risks of generative text-to-image models," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Aug. 2023, pp. 396–410.

[14] O. Parraga, M. D. More, C. M. Oliveira, N. S. Gavenski, L. S. Kupssinskü, A. Medronha, L. V. Moura, G. S. Simões, and R. C. Barros, "Fairness in deep learning: A survey on vision and language research," *ACM Comput. Surveys*, vol. 57, no. 6, pp. 1–40, Jun. 2025.

[15] P. Nemani, Y. D. Joel, P. Vijay, and F. F. Liza, "Gender bias in transformers: A comprehensive review of detection and mitigation strategies," *Natural Lang. Process. J.*, vol. 6, Mar. 2024, Art. no. 100047.

[16] S. Prerak, "Addressing bias in text-to-image generation: A review of mitigation methods," in *Proc. 3rd Int. Conf. Smart Technol. Syst. Next Gener. Comput. (ICSTSN)*, Jul. 2024, pp. 1–6.

[17] A. K. Saxena, "Quantitative measurement of bias in AI-generated content: A comprehensive narrative literature review," in *Proc. IEEE Int. Symp. Technol. Soc. (ISTAS)*, Sep. 2024, pp. 1–5.

[18] S. A. S. Williams, *Bias, Race*. Boston, MA, USA: Springer, 2011, pp. 235–237, doi: 10.1007/978-0-387-79061-9_329.

[19] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs," 2021, *arXiv:2111.02114*.

[20] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 2443–2449.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[22] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," 2018, *arXiv:1809.11096*.

[23] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.

[24] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. H. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5B: An open large-scale dataset for training next generation image-text models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 25278–25294.

[25] StabilityAI. (2022). *Stable Diffusion 2*. Accessed: Apr. 17, 2023. [Online]. Available: https://huggingface.co/stabilityai/stable-diffusion-2

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

[27] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.

[28] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3557–3567.

[29] J. Yu, Y. Xu, J. Yu Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. Karagol Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, and Y. Wu, "Scaling autoregressive models for content-rich text-to-image generation," 2022, *arXiv:2206.10789*.

[30] StabilityAI. (2023). *DeepFloyd IF: A Novel State-of-the-Art Open-Source Text-to-Image Model With a High Degree of Photorealism and Language Understanding*. Accessed: Apr. 17, 2023. [Online]. Available: https://github.com/deep-floyd/IF

[31] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, and Y. Guo, "Improving image generation with better captions," *Comput. Sci.*, vol. 2, no. 3, p. 8, 2023. [Online]. Available: https://cdn.openai.com/papers/dall-e-3

[32] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol. NAACL-HLT*, vol. 1, Jun. 2019, pp. 4171–4186.

[33] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[34] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2019.

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[36] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.

[37] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.

[38] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019.

[39] M. Razghandi, H. Zhou, M. Erol-Kantarci, and D. Turgut, "Variational autoencoder generative adversarial network for synthetic data generation in smart home," in *Proc. IEEE Int. Conf. Commun.*, May 2022, pp. 4781–4786.

[40] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," 2021, *arXiv:2112.10741*.

[41] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, and E. Raff, "VQGAN-CLIP: Open domain image generation and editing with natural language guidance," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 88–105.

[42] The Oxford Review. (Apr. 7, 2024). *The Oxford Review - Organisational & Leadership Research*. Accessed: Apr. 7, 2024. [Online]. Available: https://oxford-review.com/

[43] T. E. Yingst, *Cultural Bias*. Boston, MA, USA: Springer, 2011, p. 446, doi: 10.1007/978-0-387-79061-9_749.

[44] C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman, "Science faculty's subtle gender biases favor male students," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 41, pp. 16474–16479, 2012.

[45] Catalogue of Bias Collaboration. (2017). *Racial Bias*. Accessed: Apr. 17, 2023. [Online]. Available: https://catalogofbias.org/biases/racial-bias/

[46] J. Cevik, B. Lim, I. Seth, F. Sofiadellis, R. J. Ross, R. Cuomo, and W. M. Rozen, "Assessment of the bias of artificial intelligence generated images and large language models on their depiction of a surgeon," *ANZ J. Surg.*, vol. 94, no. 3, pp. 287–294, Mar. 2024.

[47] C. S. Webster, S. Taylor, C. Thomas, and J. M. Weller, "Social bias, discrimination and inequity in healthcare: Mechanisms, implications and recommendations," *BJA Educ.*, vol. 22, no. 4, pp. 131–137, Apr. 2022.

[48] D. E. Rupp, S. J. Vodanovich, and M. Crede, "Age bias in the workplace: The impact of ageism and causal attributions 1," *J. Appl. Social Psychol.*, vol. 36, no. 6, pp. 1337–1364, 2006.

[49] K. Crenshaw, "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics," in *Feminist Legal Theories*. Routledge, 2013, pp. 23–51.

[50] H. Ayyash-Abdo, "Status of female teachers in the middle east and North Africa region," *J. In-Service Educ.*, vol. 26, no. 1, pp. 191–207, Mar. 2000.

[51] F. van de Vijver and N. K. Tanzer, "Bias and equivalence in cross-cultural assessment: An overview," *Eur. Rev. Appl. Psychol.*, vol. 54, no. 2, pp. 119–135, Jun. 2004.

[52] S. G. Hiyasova, M. G. Mustafaeva, and F. M. Mustafaev, "Reflection of prejudices and stereotypes in cross-cultural communication," *Sci. almanac Black Sea region countries*, vol. 15, no. 3, pp. 23–29, 2018.

[53] H. Mannering, "Analysing gender bias in text-to-image models using object detection," 2023, *arXiv:2307.08025*.

[54] N. Masrourisaadat, N. Sedaghatkish, F. Sarshartehrani, and E. A. Fox, "Analyzing quality, bias, and performance in text-to-image generative models," 2024, *arXiv:2407.00138*.

[55] A. Caliskan, "Artificial intelligence, bias, and ethics," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, Aug. 2023, pp. 7007–7013.

[56] R. Ali, O. Y. Tang, I. D. Connolly, H. F. Abdulrazeq, F. N. Mirza, R. K. Lim, B. R. Johnston, M. W. Groff, T. Williamson, K. Svokos, T. J. Libby, J. H. Shin, Z. L. Gokaslan, C. E. Doberstein, J. Zou, and W. F. Asaad, "Demographic representation in 3 leading artificial intelligence text-to-image generators," *JAMA Surg.*, vol. 159, no. 1, pp. 87–95, Jan. 2024.

[57] M. Gisselbaek, M. Suppan, L. Minsart, E. Köselerli, S. Nainan Myatra, I. Matot, O. L. Barreto Chang, S. Saxena, and J. Berger-Estilita, "Representation of intensivists' race/ethnicity, sex, and age by artificial intelligence: A cross-sectional study of two text-to-image models," *Crit. Care*, vol. 28, no. 1, p. 363, Nov. 2024.

[58] H. Luo, Z. Deng, R. Chen, and Z. Liu, "FAIntbench: A holistic and precise benchmark for bias evaluation in text-to-image models," 2024, *arXiv:2405.17814*.
[59] M. D'Incà, E. Peruzzo, M. Mancini, X. Xu, H. Shi, and N. Sebe, "GradBias: Unveiling word influence on bias in text-to-image generative models," 2024, *arXiv:2408.16700*.
[60] S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite, "Stable bias: Evaluating societal representations in diffusion models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 56338–56351.
[61] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan, "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Jun. 2023, pp. 1493–1504.
[62] C. Clemmer, J. Ding, and Y. Feng, "PreciseDebias: An automatic prompt engineering approach for generative AI to mitigate image demographic biases," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 8581–8590.
[63] H. Yesiltepe, K. Akdemir, and P. Yanardag, "MIST: Mitigating intersectional bias with disentangled cross-attention editing in text-to-image diffusion models," 2024, *arXiv:2403.19738*.
[64] X. Wang, X. Yi, X. Xie, and J. Jia, "Embedding an ethical mind: Aligning text-to-image synthesis via lightweight value optimization," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 3558–3567.
[65] A. Naseh, J. Roh, E. Bagdasaryan, and A. Houmansadr, "Backdooring bias into text-to-image models," 2024, *arXiv:2406.15213*.
[66] P. P. S. Dammu, Y. Feng, and C. Shah, "Addressing weak decision boundaries in image classification by leveraging web search and generative models," 2023, *arXiv:2310.19986*.
[67] Y. Jiang, Y. Lyu, Z. He, B. Peng, and J. Dong, "Mitigating social biases in text-to-image diffusion models via linguistic-aligned attention guidance," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 3391–3400.
[68] B. Liu, L. Wang, C. Lyu, Y. Zhang, J. Su, S. Shi, and Z. Tu, "On the cultural gap in text-to-image generation," in *Proc. ECAI*, 2024, pp. 930–937.
[69] H. Luo, H. Huang, Z. Deng, X. Li, H. Wang, Y. Jin, Y. Liu, W. Xu, and Z. Liu, "BIGbench: A unified benchmark for evaluating multi-dimensional social biases in text-to-image models," 2024, *arXiv:2407.15240*.
[70] A. Sathe, P. Jain, and S. Sitaram, "A unified framework and dataset for assessing societal bias in vision-language models," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2024, pp. 1208–1249.
[71] M. Sami, A. Sami, and P. Barclay, "A case study of fairness in generated images of large language models for software engineering tasks," in *Proc. IEEE Int. Conf. Softw. Maintenance Evol. (ICSME)*, Oct. 2023, pp. 391–396.
[72] M. Flathers, G. Smith, E. Wagner, C. E. Fisher, and J. Torous, "AI depictions of psychiatric diagnoses: A preliminary study of generative image outputs in midjourney V.6 and DALL-E 3," *BMJ Mental Health*, vol. 27, no. 1, Dec. 2024, Art. no. e301298.
[73] A. Sureddy, D. Padalia, N. Periyakaruppa, O. Saha, A. Williams, A. Romero-Soriano, M. Richards, P. Kirichenko, and M. Hall, "Decomposed evaluations of geographic disparities in text-to-image models," 2024, *arXiv:2406.11988*.
[74] L. Zhang, X. Liao, Z. Yang, B. Gao, C. Wang, Q. Yang, and D. Li, "Partiality and misconception: Investigating cultural representativeness in text-to-image models," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2024, pp. 1–25.
[75] K. Ghate, A. Choudhry, and V. Bannihatti Kumar, "Evaluating gender bias in multilingual multimodal AI models: Insights from an Indian context," in *Proc. 5th Workshop Gender Bias Natural Lang. Process. (GeBNLP)*, 2024, pp. 338–350.
[76] M. Apiola, H. Vartiainen, and M. Tedre, "First year CS students exploring and identifying biases and social injustices in text-to-image generative AI," in *Proc. Innov. Technol. Comput. Sci. Educ.*, Jul. 2024, pp. 485–491.
[77] A. Chinchure, P. Shukla, G. Bhatt, K. Salij, K. Hosanagar, L. Sigal, and M. Turk, "TIBET: Identifying and evaluating biases in text-to-image generative models," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2023, pp. 429–446.
[78] M. D'Incà, E. Peruzzo, M. Mancini, D. Xu, V. Goe, X. Xu, Z. Wang, H. Shi, and N. Sebe, "OpenBias: Open-set bias detection in text-to-image generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 12225–12235.
[79] J. Cho, A. Zala, and M. Bansal, "DALL-EVAL: Probing the reasoning skills and social biases of text-to-image generation models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3043–3054.
[80] J. Wang, X. Gabby Liu, Z. Di, Y. Liu, and X. Eric Wang, "T2IAT: Measuring Valence and stereotypical biases in text-to-image generation," 2023, *arXiv:2306.00905*.
[81] P. Seshadri, S. Singh, and Y. Elazar, "The bias amplification paradox in text-to-image generation," 2023, *arXiv:2308.00755*.
[82] Kim, Y., Na, B., Park, M., Jang, J., Kim, D., Kang, W., & Moon, I. C. (2024). Training unbiased diffusion models from biased dataset. In The Twelfth International Conference on Learning Representations.
[83] G. Currie, C. Chandra, and H. Kiat, "Gender bias in text-to-image generative artificial intelligence when representing cardiologists," *Information*, vol. 15, no. 10, p. 594, Sep. 2024.
[84] A. Chauhan, T. Anand, T. Jauhari, A. Shah, R. Singh, A. Rajaram, and R. Vanga, "Identifying race and gender bias in stable diffusion AI image generation," in *Proc. IEEE 3rd Int. Conf. AI Cybersecurity (ICAIC)*, Feb. 2024, pp. 1–6.
[85] Y. Wu, Y. Shen, M. Backes, and Y. Zhang, "Image-perfect imperfections: Safety, bias, and authenticity in the shadow of text-to-image model evolution," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Dec. 2024, pp. 4837–4851.
[86] Y. Wan and K.-W. Chang, "The male CEO and the female assistant: Evaluation and mitigation of gender biases in text-to-image generation of dual subjects," 2024, *arXiv:2402.11089*.
[87] P. Wootaek Shin, J. Janice Ahn, W. Yin, J. Sampson, and V. Narayanan, "Can prompt modifiers control bias? A comparative analysis of text-to-image generative models," 2024, *arXiv:2406.05602*.
[88] Z. Ma, M. Pan, W. Wu, K. Cheng, J. Zhang, S. Huang, and J. Chen, "Food-500 cap: A fine-grained food caption benchmark for evaluating vision-language models," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 5674–5685.
[89] A. Lin, L. Monteiro Paes, S. Harsha Tanneru, S. Srinivas, and H. Lakkaraju, "Word-level explanations for analyzing bias in text-to-image models," 2023, *arXiv:2306.05500*.
[90] S. Ghosh, "Interpretations, representations, and stereotypes of caste within text-to-image generators," in *Proc. AAAI/ACM Conf. AI*, vol. 7, 2024, pp. 490–502.
[91] H. Rosenberg, S. Ahmed, G. V. Ramesh, R. K. Vinayak, and K. Fawaz, "Limitations of face image generation," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 14838–14846.
[92] T. Lee, M. Yasunaga, C. Meng, Y. Mai, J.-S. Park, A. Gupta, Y. Zhang, D. Narayanan, H. Teufel, M. Bellagente, M. Kang, T. Park, J. Leskovec, J.-Y. Zhu, L. Fei-Fei, J. Wu, S. Ermon, and P. Liang, "Holistic evaluation of text-to-image models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 69981–70011.
[93] J. Vice, N. Akhtar, R. Hartley, and A. Mian, "Quantifying bias in text-to-image generative models," 2023, *arXiv:2312.13053*.
[94] R. Naik and B. Nushi, "Social biases through the text-to-image generation lens," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Aug. 2023, pp. 786–808.
[95] A. Baines, L. Gruia, G. Collyer-Hoar, and E. Rubegni, "Playgrounds and prejudices: Exploring biases in generative AI for children," in *Proc. 23rd Annu. ACM Interact. Design Children Conf.*, Jun. 2024, pp. 839–843.
[96] M. M. Kaufenberg-Lashua, J. K. West, J. J. Kelly, and V. A. Stepanova, "What does AI think a chemist looks like? An analysis of diversity in generative AI," *J. Chem. Educ.*, vol. 101, no. 11, pp. 4704–4713, Nov. 2024.
[97] S. Kuchlous, M. Li, and J. G. Wang, "Bias begets bias: The impact of biased embeddings on diffusion models," 2024, *arXiv:2409.09569*.
[98] S. Sagar, A. Taparia, and R. Senanayake, "Failures are fated, but can be faded: Characterizing and mitigating unwanted behaviors in large-scale vision and language models," 2024, *arXiv:2406.07145*.
[99] B. Pal, A. Kannan, K. R. Prabhakar, A. J. O'Toole, and R. Chellappa, "GAMMA-FACE: Gaussian mixture models amend diffusion models for bias mitigation in face images," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2024, pp. 471–488.
[100] G. Currie, G. John, and J. Hewis, "Gender and ethnicity bias in generative artificial intelligence text-to-image depiction of pharmacists," *Int. J. Pharmacy Pract.*, vol. 32, no. 6, pp. 524–531, Nov. 2024.

[101] T. Fadahunsi, G. d'Aloisio, A. Di Marco, and F. Sarro, "How do generative models draw a software engineer? A case study on stable diffusion bias," 2025, *arXiv:2501.09014*.
[102] E. J. York, E. Brumberger, and L. V. A. Harris, "Prompting bias: Assessing representation and accuracy in AI-generated images," in *Proc. 42nd ACM Int. Conf. Design Commun.*, Oct. 2024, pp. 106–115.
[103] A. Abrar, N. Tabassum Oeshy, M. Kabir, and S. Ananiadou, "Religious bias landscape in language and text-to-image models: Analysis, detection, and debiasing strategies," 2025, *arXiv:2501.08441*.
[104] L. Sun, M. Wei, Y. Sun, Y. J. Suh, L. Shen, and S. Yang, "Smiling women pitching down: Auditing representational and presentational gender biases in image-generative AI," *J. Comput.-Mediated Commun.*, vol. 29, no. 1, p. 45, Nov. 2023.
[105] A. Jha, V. Prabhakaran, R. Denton, S. Laszlo, S. Dave, R. Qadri, C. Reddy, and S. Dev, "ViSAGe: A global-scale analysis of visual stereotypes in text-to-image generation," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics*, 2024, pp. 12333–12347.
[106] P. Esposito, P. Atighehchian, A. Germanidis, and D. Ghadiyaram, "Mitigating stereotypical biases in text to image generative systems," 2023, *arXiv:2310.06904*.
[107] M. Erfanian, H. V. Jagadish, and A. Asudeh, "Chameleon: Foundation models for fairness-aware multi-modal data augmentation to enhance coverage of minorities," 2024, *arXiv:2402.01071*.
[108] M. Hou, Y. Wu, C. Xu, Y.-H. Huang, C. Bai, L. Wu, and J. Bian, "InvDiff: Invariant guidance for bias mitigation in diffusion models," 2024, *arXiv:2412.08480*.
[109] G. Wang and L. Specia, "MoESD: Mixture of experts stable diffusion to mitigate gender bias," 2024, *arXiv:2407.11002*.
[110] B. Krojer, E. Poole-Dayan, V. Voleti, C. Pal, and S. Reddy, "Are diffusion models vision-and-language reasoners?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 8385–8405.
[111] A. Arias-Duart, V. Gimenez-Abalos, U. Cortés, and D. Garcia-Gasulla, "Assessing biases through visual contexts," *Electronics*, vol. 12, no. 14, p. 3066, Jul. 2023.
[112] M. M. Tanjim, K. K. Singh, K. Kafle, R. Sinha, and G. W. Cottrell, "Discovering and mitigating biases in CLIP-based image editing," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 2972–2981.
[113] G. Currie, J. Currie, S. Anderson, and J. Hewis, "Gender bias in generative artificial intelligence text-to-image depiction of medical students," *Health Educ. J.*, vol. 83, no. 7, pp. 732–746, Nov. 2024.
[114] S. Wang, X. Cao, J. Zhang, Z. Yuan, S. Shan, X. Chen, and W. Gao, "VLBiasBench: A comprehensive benchmark for evaluating bias in large vision-language model," 2024, *arXiv:2406.14194*.
[115] W. ELsharif, M. Agus, M. Alzubaidi, and J. She, "Cultural relevance index: Measuring cultural relevance in AI-generated images," in *Proc. IEEE 7th Int. Conf. Multimedia Inf. Process. Retr. (MIPR)*, vol. 33, Aug. 2024, pp. 410–416.

WALA ELSHARIF received the bachelor's degree in information technology from the University of Khartoum and the master's degree in computer science from The University of Texas at San Antonio. She is currently pursuing the Ph.D. degree in computer science with Hamad Bin Khalifa University.

Her research interests include generative AI, human–computer interaction, and cultural studies, exploring the intersection of artificial intelligence and societal contexts. Her work examines the cultural relevance of AI-generated content and the ways AI can be adapted for diverse cultural settings. She has contributed to research on cultural representation in TTI models and the development of evaluation metrics for AI-generated images.

MAHMOOD ALZUBAIDI received the master's degree in internet engineering from the National Advanced IPv6 Center, Universiti Sains Malaysia, in 2018, and the Ph.D. degree from Hamad Bin Khalifa University, Qatar, in 2023, where he continues to contribute to the field as a Researcher. His research interests are broad, spanning across the Internet of Things (IoT), deep learning, machine learning, image segmentation, generative AI, and AI application in healthcare.

MARCO AGUS (Member, IEEE) received the M.Sc. and Ph.D. degrees from the University of Cagliari, Italy.

He was a Research Engineer with the King Abdullah University of Science and Technology, Saudi Arabia, and a Research Scientist with the Center of Research, Development, and Advanced Studies (CRS4), Cagliari, Italy. He is currently an Associate Professor with the College of Science and Engineering, Hamad Bin Khalifa University. His research interests span different domains in visual computing, from haptics and visual rendering for medical applications to real-time exploration of massive models, to machine learning methods for electron microscopy biology data, and indoor environments. He taught courses at several important visual computing venues, including IEEE CVPR, ACM SIGGRAPH, and Eurographics, and he regularly acts as a committee member, a reviewer, the chair, and an associate editor of top journals and conferences in the visual computing domain.

Open Access funding provided by 'Qatar National Library' within the CRUI CARE Agreement