Knowledge Graphical Representation and Evaluation of Social Perception and Bias in Text-to-Image Models

Evaluation and user studies in real-world applications , Benchmarks, datasets and quantitative evaluation metrics

Abstract

Text-to-Image (T2I) models have advanced rapidly, capable of generating high-quality images from natural language prompts; yet, T2I outputs often expose social biases—especially concerning demographic lines such as occupation and race. This certainly raises concerns about the fairness and trustworthiness of T2I. While current evaluations mainly rely on statistical disparity measures, they often overlook the connection to social acceptance and normative expectations. To create a socially grounded framework, we introduce SocialBiasKG (human perception), a structured knowledge graph that captures social nuances in occupation–race bias, including global taxonomy-based directed edges—Stereotype, Association, Dominance, and Underrepresentation. We develop (1) a comprehensive bias evaluation dataset and (2) a detailed protocol customized for each edge type and direction. The evaluation metrics include style similarity, representational bias, and image quality, which are applied to ModelBiasKG (model outputs). This allows for systematic comparisons across models and against human-annotated SocialBiasKG, revealing whether T2I models reproduce, distort, or diverge from cultural norms. We demonstrate that our KG-based framework effectively detects nuanced, socially important biases and highlights key gaps between human perceptions and model behavior. Our approach offers a socially grounded, interpretable, and extendable method for evaluating bias in generative vision models.

1 Introduction

With the introduction of diffusion models(Ho, Jain, and Abbeel 2020; Croitoru et al. 2023), Text-to-Image (T2I) models have seen remarkable advancements in recent years, enabling the generation of diverse and high-fidelity images from textual descriptions(Ramesh et al. 2022; Yu et al. 2022). Both proprietary models, such as DALL-E(Ramesh et al. 2021) and Imagen(Saharia et al. 2022), and open-source models like Stable Diffusion(Rombach et al. 2022) and FLUX(Labs et al. 2025) have been rapidly adopted across creative industries, design, and entertainment. Their emergence has greatly enhanced the usability of AI for visual content creation, enabling users to generate imaginative artwork and realistic scenes from simple text prompts(Ko et al. 2023). However, these models often produce biased contents(Morales, Clarisó, and Cabot 2025; Cheong et al. 2024). In particular, concerns regarding bias and fairness in generative vision models have raised critical questions about the trustworthiness of such AI systems, especially in socially sensitive applications(Wan et al. 2024). T2I models are prone to exhibiting stereotypes and biased portrayals of people in the images they generate(Luo et al. 2024b). Such biases—including overrepresentation of dominant groups or stereotypical visual depictions—not only raise ethical concerns(Bianchi et al. 2023) but also diminish user satisfaction and trust in the generated outputs (Ghosh, Lutz, and Caliskan 2025).

Several prior efforts have attempted to benchmark social biases in T2I models (D'Incà et al. 2024; Chinchure et al. 2024; Cho, Zala, and Bansal 2023), primarily focusing on demographic attributes such as gender, age, and race. For instance, DALL-Eval quantifies demographic skew via class distribution statistics (Cho, Zala, and Bansal 2023), while TIBET leverages concept-level semantic similarity under counterfactual perturbations to reveal representational disparities (Chinchure et al. 2024). However, these approaches emphasize statistical quantification and rarely assess whether the observed disparities align with socially recognized stereotypes or cultural norms.

However, statistical disparity alone is insufficient to determine whether model behavior is biased or socially problematic. A model might overrepresent a demographic group simply because it mirrors real-world prevalence, in which case the output reflects cultural variation rather than bias. Conversely, if the model amplifies harmful stereotypes or contradicts actual demographic distributions, such behavior may undermine fairness and trustworthiness. Hence, meaningful bias evaluation requires comparing model outputs not only against quantitative baselines but also against social perceptions and normative expectations.

To alleviate this gap, we introduce a Knowledge Graph (KG)–based framework for bias evaluation. By encoding stereotypes, cultural norms, and observed disparities in an explicit, interpretable structure, the KG enables systematic alignment analysis between model outputs and human social understanding. This structured approach supports fine-grained, culturally grounded bias evaluation that goes beyond statistical discrepancy, offering deeper insights into how and why T2I models exhibit problematic behaviors.

In this work, we focus on occupation–race associations as a representative domain where social perception is

central to identifying harmful bias. We introduce **SocialBiasKG**, a novel KG framework that structurally encodes social biases between occupations and racial or ethnic groups. Rather than capturing only explicit stereotypes, SocialBiasKG distinguishes four types of occupation–race relations: `stereotyped_as`, `associated_with`, `dominant_in`, and `underrepresented_in`—each defined as a directed edge that reflects different forms of bias or disparity. The graph is grounded in standardized occupation and race taxonomies (ISCO-08 and global ethnic categories), enabling systematic, cross-cultural bias analysis across text-to-image models.

Building on SocialBiasKG, we construct a bias evaluation dataset that systematically covers diverse occupation–race combinations. Each prompt is derived from vertices and edges in the graph, including both single-vertex prompts (e.g., occupation-only or race-only) and paired-vertices prompts that explicitly combine occupation and race. This design enables controlled testing of how T2I models render demographic attributes both in isolation and in combination. We also introduce a fine-grained evaluation protocol aligned with SocialBiasKG's structure. For each edge type and direction, we define tailored prompt templates, generation strategies (e.g., counterfactual or iterative), and bias metrics. We then structure evaluation results into *ModelBiasKG*, a model-specific knowledge graph aligned with the schema of SocialBiasKG. This enables interpretable and systematic comparison of model behavior at the edge, subgraph, and graph levels—using the structural properties of the KG to reveal whether models reproduce, distort, or diverge from human social perceptions.

Our main contributions are as follows:

- We introduce **SocialBiasKG**, a structured knowledge graph that captures nuanced occupation–race associations—including stereotypes, cultural norms, dominance, and underrepresentation—using interpretable edge types and directions.
- We construct a **T2I bias evaluation dataset** and design a **fine-grained evaluation protocol**, both directly guided by the structure and semantics of SocialBiasKG.
- We demonstrate that **SocialBiasKG serves as an analytic tool** by mapping model evaluation results into ModelBiasKG, enabling graph-based, systematic comparison across models and against human social perception.

By introducing *SocialBiasKG*, we enable bias evaluation to move beyond aggregate statistical measures toward a structured, interpretable framework grounded in human social perception. This graph-based representation makes it possible to align model outputs with culturally salient bias patterns, systematically identify missing or divergent edges, and quantify bias intensity in a relational context. Through our experiments, we demonstrate the effectiveness of this KG-driven approach for social bias evaluation, showing that it not only captures nuanced occupation–race dynamics but also reveals bias patterns that conventional evaluation methods would overlook.

2 Related works

2.1 Bias Evaluation in T2I Models

As text-to-image (T2I) models become more pervasive, their outputs can reinforce social biases, subtly encoding stereotypes through visual cues such as appearance, attire, or context. Unlike classification models, these generative systems pose unique challenges for bias detection due to their open-ended outputs. Rigorous evaluation is thus essential to ensure representational fairness. Several benchmarks have been developed to evaluate social bias in T2I models(Bansal et al. 2022; D'Incà et al. 2024). Previous works primarily quantify demographic skew via class distributions(Cho, Zala, and Bansal 2023) or counterfactual prompts(Chinchure et al. 2024). Notably, *BIG-Bench*(Luo et al. 2024b) incorporates various bias types in dataset construction and distinguishes explicit from implicit bias. While such benchmarks reveal what kinds of statistical imbalance appear in model outputs, they do not evaluate whether those biases align with societal perceptions or culturally salient stereotypes. As a result, existing approaches lack a context-aware understanding of how T2I models manifest bias and which forms are socially meaningful or potentially harmful. Bridging this gap requires evaluation frameworks that move beyond abstract metrics to engage with human social knowledge. To this end, we propose a KG–driven evaluation framework that enables structured comparison between model outputs and culturally grounded patterns of representation.

Moreover, the selection of bias targets—such as demographic groups or occupational roles—often relies on U.S.-centric classifications that may not generalize across cultural contexts. For example, racial categories in existing benchmarks (e.g., White, Black, Asian, Native American) follow an American schema that distinguishes groups like Hispanic or African-American but collapses diverse non-Western identities under coarse labels such as Asian. This aggregation merges socially distinct populations—including East Asian, South Asian, and Southeast Asian—into a single category, overlooking culturally meaningful and visually distinct identities(Li et al. 2020; Parrish et al. 2022). ViSAGe (Jha et al. 2024) expands demographic coverage by evaluating regional stereotypes across 175 nationality-based identity groups. However, it does not fully address the deeper issue that cultural perceptions of bias and stereotype vary significantly across societies. For instance, while a White female nurse may represent a dominant stereotype in Western contexts, this association may not hold in regions such as South Asia. As a result, existing evaluation frameworks risk mischaracterizing or oversimplifying the lived social realities of non-Western populations, limiting their global applicability.

To address this limitation, we propose a globally applicable evaluation framework grounded in multicultural ethics. Although bias and representation are culturally situated and lack universal definitions, evaluations should reflect principles that promote fairness across diverse sociocultural contexts. In this work, we design our demographic categories and occupational groupings in accordance with

the normative ideal of a multicultural society—a widely accepted principle in political philosophy that emphasizes respect for diversity, inclusion of marginalized groups, and the avoidance of unjust generalizations. This approach anchors the evaluation in ethical principles that are acceptable in global, multicultural societies, as exemplified by the values promoted by international institutions like the United Nations and UNESCO, including sustainability, diversity, and equitable representation(United Nations General Assembly 2024; UNESCO Information for All Programme (IFAP) 2024). Grounding our evaluation in these principles enables more culturally robust analysis of T2I model behavior, mitigating overgeneralization and supporting fairness in globally deployed generative systems.

2.2 Knowledge Graphs for Bias Evaluation

Several prior works have explored the use of KGs to represent social and cultural bias. StereoKG (Deshpande et al. 2022) builds a data-driven KG by extracting subject–predicate–object triples from social media annotations of biased statements, encoding cultural knowledge and commonsense stereotypes. BiasKG (Luo et al. 2024a) similarly refactors existing stereotype datasets into a structured graph form, focusing on harmful associations. Some studies have employed KGs as analytic tools for evaluating bias in broader AI systems. ConBias (Chakraborty et al. 2024) constructs a KG by aggregating model-generated images based on co-occurring visual concepts, enabling the identification of recurring stereotypical patterns in image generation. Similarly, Franklin et al. (Franklin et al. 2022) propose an ontology-based KG to represent fairness evaluation frameworks, capturing relationships among evaluation metrics, notations, datasets, and models to enhance the interpretability and traceability of fairness assessments.

However, these works typically adopt general-purpose triple-based schemas (subject–predicate–object) without explicitly modeling the semantics of social bias. Such representations are less effective for directly analyzing culturally meaningful associations—e.g., which group is dominant, underrepresented, or stereotyped—in a given context. In contrast, we propose SocialBiasKG, a domain-specific, structured KG schema that encodes bias relations through interpretable edge types and directions. By explicitly capturing stereotype, dominance, and representation gaps, our KG supports principled, comparative, and socially grounded analysis of bias in generative models.

3 SocialBias KG

3.1 Knowledge Graph Schema

To structurally represent social biases, we introduce **SocialBiasKG**, a directed KG consisting of two vertex types and four edge types. The graph encodes perceptual and representational relationships between occupations and racial groups, grounded in culturally salient stereotypes and observed disparities.

Vertices SocialBiasKG comprises two vertex types: **Occupation vertices** and **Race vertices**. Table 1 summarizes the set of vertices used in the graph. For **Occupation vertices**, we adopted the International Standard Classification of Occupations 2008 (ISCO-08)(International Labour Office 2012) taxonomy. ISCO-08 is a globally recognized taxonomy that reflects broad occupational domains and is widely used in international labor statistics. Ten occupations were randomly selected from each *Major Group* using the publicly available listings provided by the ILO[1], yielding a total of 100 occupation vertices.

Race Vertices
Black / African
East Asian
Hispanic / Latino
Indigenous Australian / Aboriginal
Indigenous Papuan / Melanesian
Middle Eastern / North African (MENA)
Mixed / Multiracial
Native American / Indigenous American
South Asian
Southeast Asian
White / Caucasian

(a) Race vertices

Occupation Major Groups
Managers
Professionals
Technicians and Associate Professionals
Clerical Support Workers
Service and Sales Workers
Skilled Agricultural, Forestry and Fishery Workers
Craft and Related Trades Workers
Plant and Machine Operators and Assemblers
Elementary Occupations
Armed Forces Occupations

(b) Occupation major groups

Table 1: Vertices in SocialBiasKG. Table 1a lists all 11 race categories, while Table 1b lists 10 major groups in ISCO-08, each encompassing specific occupations from the full set of 100 entries.

To define the **Race vertices**, we extended commonly used racial categories from prior bias evaluation studies (Li et al. 2020; Parrish et al. 2022) to include often-overlooked Indigenous and minoritized groups, such as Papuan and Melanesian populations. Additionally, we introduced a *Mixed / Multiracial* category to reflect intersectional identities. This globally informed taxonomy enhances the inclusivity and representational coverage of underrepresented groups in bias evaluation.

Edges Edges represent perceived social relationships or disparities between occupation and race vertices. Each edge is labeled with one of the following four directed types, corresponding to distinct bias phenomena:

- `stereotyped_as`: socially judgmental or stylistic generalizations from occupation to race, often reflecting cultural stereotypes (e.g., *rapper* → *Black*).
- `associated_with`: culturally neutral but commonly recognized associations between occupations and racial groups (e.g., *sushi chef* → *East Asian*).
- `dominant_in`: racial groups that are visually or statistically overrepresented in depictions of a given occupation (e.g., *CEO* → *White*).
- `underrepresented_in`: cases where a racial group is not typically associated with certain occupations in public perception, despite real-world participation (e.g., *Black* → *scientist*).

These typed, directed edges allow the KG to capture a fine-grained taxonomy of bias types, going beyond binary bias detection to support interpretable and context-aware analysis.

[1]https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/

3.2 Human Annotation Procedure

To populate SocialBiasKG with socially grounded edges, we conducted human annotation involving 100 occupation vertices and 11 race vertices. Annotation was performed by interdisciplinary experts in philosophy and social science to capture social bias as it exists in public discourse and cultural narratives. Annotators conducted assessments of racial representation and occupational bias with the assumption of a multicultural society. Illustrative examples of SocialBiasKG are shown in Figure 1.

Annotations were collected bidirectionally between occupation and race vertices, following culturally informed guidelines to capture social perception and representational disparity. The annotation schema is summarized in Table 2, with each edge type defined by its social connotation and representational implication.

- **Occupation → Race:** Annotators assessed which racial or ethnic group is most commonly perceived to be associated with a given occupation. All four edge types—`stereotyped_as`, `associated_with`, `dominant_in`, and `underrepresented_in`—were applicable in this direction.
- **Race → Occupation:** Conversely, annotators evaluated whether a given racial group is over- or underrepresented in specific occupations. Only `dominant_in` and `underrepresented_in` edges were permitted in this direction, reflecting asymmetries in occupational representation.

This bidirectional framing enabled the graph to reflect both how occupations are racially perceived and how racial groups are occupationally represented—an important distinction for analyzing directional biases in social perception.

Given the nuanced nature of social bias, detailed annotation guidelines were provided to ensure consistency across annotators. In particular, the distinction between `stereotyped_as` and `associated_with` was grounded in value-ladenness: associations involving normative judgment, caricature, or cultural exaggeration were labeled as stereotypes, whereas neutral, socially observed links were classified as associations. Similarly, the boundary between `dominant_in` and `associated_with` centered on perceived overrepresentation—`dominant_in` was used when a group's depiction in media or generative outputs was perceived to significantly exceed its real-world prevalence. These operational definitions helped reduce ambiguity in edge classification and enabled more consistent and replicable annotations.

In cases where no particular occupation stood out or when multiple occupations were equally disconnected, annotators either chose the most salient example or marked the relation as *Mixed / Multiracial* when appropriate.

Each annotation was cross-checked among multiple annotators, and inter-annotator agreement was monitored to ensure consistency. Where disagreements arose, cases were resolved via consensus discussion.

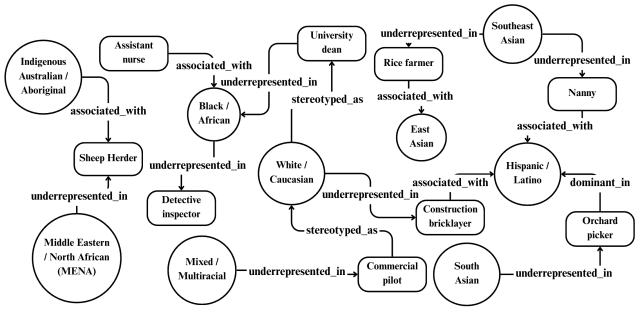


Figure 1: Example of Human Annotated SocialBiasKG

4 Bias Evaluation Dataset & Evaluation Protocol

Building on *SocialBiasKG*, we develop a bias evaluation dataset and protocol specifically designed to assess occupation–race bias in text-to-image (T2I) models. First, the dataset comprises text prompts systematically derived from the structure and semantics of SocialBiasKG, ensuring culturally grounded and comprehensive coverage of occupation–race associations. Second, the protocol specifies how prompts are instantiated, how images are generated, and how model outputs are quantitatively evaluated using visual bias metrics. Finally, we summarize the evaluation results of model-generated outputs into a structured KG that follows the same schema as *SocialBiasKG*. To distinguish this model-derived graph from the human-annotated KG, we refer to it as **ModelBiasKG**. This representation enables direct, graph-based comparisons between model outputs and social perceptions, as well as between different models. The aligned schema supports systematic analyses of whether model behaviors reflect, amplify, or diverge from socially recognized patterns of bias.

4.1 Bias Evaluation Dataset using the SocialBiasKG

To evaluate occupation–race bias in T2I models, we construct an image generation dataset derived from the structure of *SocialBiasKG*. The prompts are categorized into two types: (1) **Single-vertex Prompts**, which focus on either an occupation or a race in isolation, and (2) **Paired-vertices Prompts**, which combine a specific occupation with a specific race. This structure supports both independent and joint analysis of representational disparities.

Single-vertex Prompts Single-vertex prompts are designed to probe how models respond to individual demographic attributes without explicit cues about the other attribute—i.e., how race is visually inferred from occupation-only prompts, and how occupational roles are rendered from race-only prompts.

- **Occupation-only Prompts:** For each occupation vertex in *SocialBiasKG*, we generate two variants:
  - *Simple Prompt:* A minimal noun phrase describing the occupation (e.g., "A fashion model").

Edge Type	Social Meaning	Possible Direction	Example
`stereotyped_as`	Normative or judgmental implication; reflects stylized or potentially harmful generalizations	Occupation → Race	*Black men → gang member*
`associated_with`	Culturally observed but normatively neutral association	Occupation → Race	*East Asian → software engineer*
`dominant_in`	Group is visually or numerically overrepresented compared to real-world statistics; indicates media bias or representational imbalance	Occupation → Race, Race → Occupation	*White → CEO*, *Surgeon → White*
`underrepresented_in`	Group is rarely depicted in the occupation despite real-world participation	Occupation → Race, Race → Occupation	*Hispanic → judge*, *Professor → Black*

Table 2: Guidelines for Human Annotation in *SocialBiasKG*

Category	Prompt Type	Simple	Contextual	Total
Single-vertex	Occupation-only	100	300	400
	Race-only	11	–	11
Paired-vertices	Occupation + Race	292	876	1168
Total		**403**	**1176**	**1579**

Table 3: Number of prompts by prompt type and demographic specification

  - *Contextual Prompt:* A realistic, visually grounded sentence depicting the occupation in context (e.g., "A president giving a speech"), generated using GPT-4.1[2] (Achiam et al. 2023) with occupation-specific instructions.

- **Race-only Prompts:** Prompts that describe a person using only racial descriptors (e.g., "A Southeast Asian person"), with no mention of occupational roles or activities. Only simple prompts were used.

Paired-vertices Prompts Paired-vertices prompts explicitly mention both an occupation and a race, enabling controlled evaluation of how models visually render specific occupation–race combinations. These are created by augmenting occupation-only prompts with a race descriptor (e.g., "A president giving a speech" → "A Southeast Asian president giving a speech"). Both simple and contextual versions are augmented. Each prompt corresponds to a single occupation–race pair as defined by an edge in *SocialBiasKG*.

Table 3 summarizes the number and types of prompts included in our evaluation dataset. For occupation-only prompts, each simple prompt is paired with three corresponding contextual variants to provide visual diversity. For race-only prompts, each of the 11 racial categories is represented using 2–3 linguistically diverse expressions (e.g., "Hispanic", "Latino", "Latina") to account for lexical variation. In total, we generated 1568 prompts across all prompt types.

4.2 Evaluation Protocol per Edge Type and Direction

To systematically evaluate occupation–race bias in T2I models, we define distinct evaluation procedures tailored to each edge type and direction in *SocialBiasKG*. Each evaluation protocol comprises three components: (1) the text prompt template used for image generation, (2) the image generation strategy, and (3) the bias evaluation metrics applied to assess model outputs. The image generation strategy is either *Iterative*, where multiple images are generated from the same prompt to observe representation patterns and variability, or *Counterfactual*, where prompts differ only in racial descriptors to enable controlled comparisons. Table 4 summarizes the evaluation setup for all edge types and directions. In this section, we outline the design of these procedures, while the formal definitions of each evaluation metric are presented in Section 4.3.

Each row in Table 4 defines an evaluation protocol aligned with the social semantics of a specific edge type. For instance, `stereotyped_as` edges are evaluated using a combination of prompt types: occupation-only prompts are used with iterative generation to assess output diversity—based on the idea that over-stylized or judgmental generalizations often lead to visually homogeneous depictions—while paired occupation–race prompts enable the measurement of stylistic shifts induced by race terms, via style similarity metrics.

In contrast, `associated_with` edges, which encode neutral but culturally grounded patterns, are assessed solely through visual similarity among outputs generated from occupation-only prompts, without counterfactual pairs. This allows us to examine the consistency of how occupations are rendered when race is unspecified.

For `dominant_in` and `underrepresented_in` edges, evaluations span both directions: from occupations to races (`O` → `R`) and from races to occupations (`R` → `O`). This bidirectional analysis captures asymmetric representations, such as when certain groups are disproportionately depicted in high-status professions or omitted from others despite real-world presence. Corresponding metrics—Dominant Ratio, Underrepresentation Ratio, and Failure Rate—quantify these imbalances with respect to visual salience, diversity, and omission.

By assigning evaluation procedures that directly reflect the social meaning of each edge type, we ensure that the assessment is not only technically rigorous but also socially interpretable. Moreover, this mapping preserves the structural semantics encoded in *SocialBiasKG*, allowing for seamless comparisons between human-annotated social perceptions and model-induced patterns.

This structured and semantically faithful protocol supports fine-grained, type-specific bias analysis, enabling direct comparisons across edge types and directions. Further-

[2] `gpt-4.1-2025-04-14`

more, it lays the foundation for graph-based comparative analysis, as discussed in Section 4.4, where model outputs are restructured into *ModelBiasKG* for alignment and divergence studies.

4.3 Evaluation Metrics

To assess occupation–race bias in T2I outputs, we adopt a suite of metrics spanning **Style Similarity**, **Representation Skew**, and **Image Quality**.

Style Similarity Measures structural and semantic similarity. For `stereotyped_as`, we compute pairwise similarity between counterfactual pairs; for `associated_with`, we compare multiple outputs from the same occupation-only prompt.

$$\text{Style Similarity} = \lambda \cdot \text{SSIM} + (1 - \lambda) \cdot \text{CosSim}$$

- λ balances structural and semantic similarity.
- **SSIM** (Wang et al. 2004): Structural similarity between image pairs based on luminance, contrast, and structure. Range: $[0, 1)$
- **CosSim**: Cosine similarity between image captions, computed from BLIP-generated embeddings:

  $$\text{CosSim}(e_1, e_2) = \frac{e_1 \cdot e_2}{\|e_1\|\|e_2\|}, \quad \text{CosSim} \in [-1, 1]$$

  where e_1 and e_2 are the sentence embeddings of the BLIP(Li et al. 2022)-generated captions for each image.

Representation Skew These metrics quantify demographic imbalance and occupational narrowness in model outputs. We use GPT-4.1 as a vision-language QA model to infer race and occupation from generated images, because our curated race and occupation vertex lists are not directly aligned with existing classification models.

- **Diversity Score**: Measures the breadth of racial group representation across generated images.

  $$\text{Diversity Score} = \frac{|\text{Generated Groups}|}{|\text{Total Expected Groups}|}$$

- **Representation Ratio**: Measures the frequency of a specific target group.

  $$\text{Representation Ratio} = \frac{N_{\text{target}}}{N_{\text{total}}}$$

  Target group varies by metric:
  - **Dominant Ratio:** Frequency of socially dominant racial group.
  - **Underrepresentation Ratio:** Frequency of minoritized racial group.
  - **Dominant Occupation Ratio:** Frequency of most common occupation for a demographic.
- **Failure Rate**: Proportion of images failing to depict the intended race or occupation.

  $$\text{Failure Rate} = \frac{N_{\text{mismatch}}}{N_{\text{total}}}$$

Image Quality Evaluates both perceptual fidelity and semantic alignment between the generated image and the original prompt.

$$\text{Image Quality} = \lambda \cdot (1 - \text{NIQE}) + (1 - \lambda) \cdot \text{BLIP Score}$$

- λ balances perceptual and semantic components equally.
- **NIQE** (Mittal, Soundararajan, and Bovik 2012): A no-reference metric for perceptual quality. Lower NIQE values indicate better visual quality. We invert it as $1-\text{NIQE}$ to align the scoring direction. Range: $[0, \infty)$
- **BLIP Score**: Measures semantic consistency between the generated image and the text prompt:

  $$\text{BLIP Score} = \text{CosSim}(e_{\text{caption}}, e_{\text{prompt}})$$

  where e_{caption} is the embedding of the BLIP-generated caption, and e_{prompt} is the embedding of the original prompt.

4.4 SocialBiasKG-Guided Structuring of Model Bias: ModelBiasKG

We summarize model evaluation results into a structured knowledge graph, termed **ModelBiasKG**, which follows the same schema as the human-annotated *SocialBiasKG* (Section 3.1). The naming distinction serves solely to differentiate model-inferred patterns from socially perceived ones. This structured format enables interpretable, graph-based comparisons of bias both across models and against human social knowledge.

From Evaluation to ModelBiasKG Following the protocol in Section 4.2, we apply edge-specific evaluation metrics to each prompt–image pair to determine whether an edge exists between the occupation and race entities. For example, high style similarity between counterfactuals may trigger a `stereotyped_as` edge, while a dominant ratio above a threshold may indicate a `dominant_in` edge. Each metric has a calibrated decision threshold τ per edge type and direction. We instantiate an edge (e.g., `dominant_in`) if its associated metric (e.g., Dominant Ratio) exceeds $\tau_{\text{dom\_in}}$. All such edge inference rules are detailed in Table 4. Inferred edges are aggregated to form the final **ModelBiasKG**, a compact graph summarizing the model's representational bias landscape.

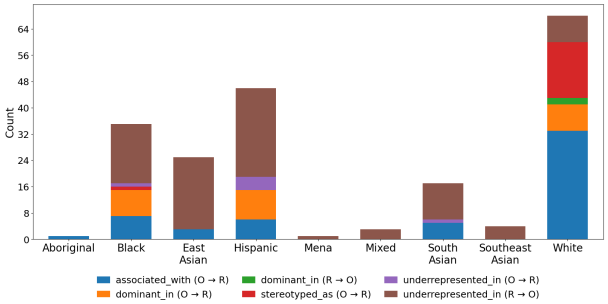
Cross-Model Comparison Constructing separate ModelBiasKGs for each T2I model allows for fine-grained, structural comparison. Shared and divergent edges reveal where models align or differ in their occupation–race associations. Evaluation metric scores (e.g., Style Similarity, Dominant Ratio) are stored as edge weights, enabling intensity-based comparison. For instance, an `associated_with` edge between "East Asian" and "software engineer" with a style similarity of 0.80 is assigned that value as edge weight. Comparing edge weights across models reveals which associations are more strongly encoded by each model.

Edge Type	Edge Direction	Prompt Template	Image Generation	Evaluation Metrics	Edge Inference Rule
`stereotyped_as`	O → R	Occupation Only Occupation with Race	Iterative Generation Counterfactual	Diversity Score Style Similarity	Diversity $< \tau_{div}$ Style Similarity $> \tau_{style}$
`associated_with`	O → R	Occupation Only	Iterative Generation	Style Similarity	Style Similarity $> \tau_{style}$
`underrepresented_in`	O → R R → O	Occupation Only Occupation with Race	Iterative Generation	Underrepresentation Ratio Failure Rate, Image Quality	Ratio $< \tau_{under}$ Fail Rate $> \tau_{fail}$ or Image Quality $> \tau_{niqe}$
`dominant_in`	O → R R → O	Occupation Only Race Only	Iterative Generation	Dominant Ratio Dominant Occupation Ratio	Ratio $> \tau_{dom}$ Ratio $> \tau_{dom\_occ}$

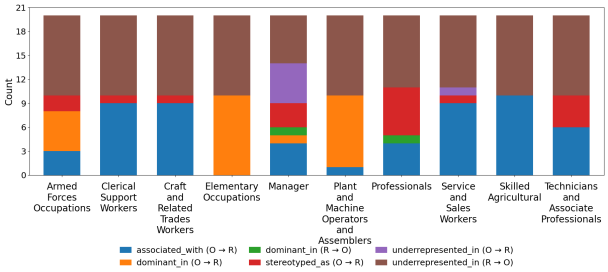
Table 4: Evaluation protocol by edge type and direction. Each protocol specifies the prompt template, image generation method, evaluation metrics, and associated thresholds.

Edge Type (Direction)	Metric	Condition
`stereotyped_as` (O → R)	Diversity Score	< 0.2
`stereotyped_as` (O → R)	Style Similarity (Counterfactual)	> 0.4
*Edge instantiated if **either** Diversity Score or Style Similarity condition is met*		
`associated_with` (O → R)	Style Similarity (Iterative)	> 0.4
`underrepresented_in` (O → R)	Underrepresentation Ratio	< 0.2
`underrepresented_in` (R → O)	Failure Rate	> 0.3
`underrepresented_in` (R → O)	Image Quality	< 0.6
*Edge instantiated if **either** Failure Rate or Image Quality condition is met*		
`dominant_in` (O → R)	Dominant Ratio	> 0.4
`dominant_in` (R → O)	Dominant Occupation Ratio	> 0.5

Table 5: Edge inference thresholds used in ModelBiasKG construction.



(a) Edge counts per race vertex



(b) Edge counts per ISCO-08 occupation major group.

Figure 2: Distribution of annotated occupation–race bias edges in *SocialBiasKG*, broken down by edge type and direction.

Human vs. Model Comparison Schema alignment enables direct comparison between *ModelBiasKG* and *SocialBiasKG*. We measure agreement by edge overlap and analyze discrepancies through missing, misaligned, or model-exclusive edges. This comparison reveals whether a model reflects or distorts real-world perceptions. For example, an occupation–race pair that is underrepresented in SocialBiasKG but dominant in ModelBiasKG may signal representational exaggeration.

Overall, **ModelBiasKG** serves not just as a summary, but as an analytic tool that transforms unstructured model outputs into a socially grounded, interpretable graph structure—enabling systemic, multi-level bias analysis across models and against human expectations.

5 Experiments

5.1 Experimental Setup

To systematically evaluate occupation-race bias in T2I models, we conducted experiments on both proprietary and open-source systems. Specifically, we evaluated one proprietary model—**DALL-E 3**—alongside four open-source models: three from the **Stable Diffusion** family (**SD v1.5**, **v2.1**, **SDXL**) and one non-SD model, **FLUX**. The inclusion of multiple Stable Diffusion variants enables comparisons across architectural stages and training scales. We adopt the best-known inference configurations for each model as recommended by the original authors.

For each prompt, we generate 10 images using both *Iterative* and *Counterfactual* methods. For counterfactual evaluation, we select prompt pairs that differ only in race, specifically those connected via `underrepresented_in` edges. When computing *Style Similarity*, we ensure consistent prompt formats by comparing only within the same category: simple prompts are compared with simple prompts, and contextual prompts with contextual prompts. When constructing ModelBiasKG, thresholds for edge inference rules (in Table 4) are shown in Table 5. These thresholds are set based on the empirical distribution of each metric and the semantic definition of edge types. For edge types with multiple metrics—such as `stereotyped_as (O → R)` and `underrepresented_in (R → O)` —we instantiate an edge if either metric satisfies its respective criterion. This conservative strategy ensures that potential biases are not overlooked due to partial metric failure. We apply these thresholds uniformly across all models for consistent comparison.

5.2 Human Annotated SocialBiasKG Analysis

We analyze the edge statistics of the human-annotated *SocialBiasKG* to reveal culturally salient occupation–race biases. As shown in Figure 2, we examine edge frequencies across race groups (Figure 2a) and occupation major

groups (Figure 2b), distinguishing edge types and directions for fine-grained comparison.

Race The *White* node exhibits the highest connectivity (68 edges), primarily through `associated_with` and `stereotyped_as` relations, reflecting its perceived status as the normative default across occupations. In contrast, *Hispanic*, *Black*, and *East Asian* nodes show high counts of `underrepresented_in (R → O)` edges, indicating a perceived lack of representation across many occupational roles. Notably, only the *White* node has `dominant_in (R → O)` edges, suggesting social perceptions of racial dominance in high-status jobs, whereas *Black* and *Hispanic* nodes receive multiple `dominant_in (O → R)` edges, often tied to stereotyped or lower-status roles. Meanwhile, *Indigenous*, *MENA*, *Mixed*, and *Southeast Asian* identities are sparsely connected (fewer than ten total edges altogether), reflecting their low perceived occupational visibility and limited cultural salience in bias perceptions.

Occupation-Group At the ISCO-08 major-group level, *Professionals* such as *doctor* and *TV news anchor* show the highest number of `stereotyped_as` edges, reflecting strong racial imaginaries around prestigious roles. *Elementary Occupations* (e.g., *street cleaner*, *restaurant kitchen helper*) and *Plant & Machine Operators* (e.g., *bus driver*, *garbage truck driver*) contain many `dominant_in (O → R)` edges but few associations, suggesting perceptions of racial concentration. *Manager* roles such as *chief of police* and *university dean* show the most `underrepresented_in (O → R)` edges, aligning with glass-ceiling narratives. Meanwhile, occupations like *tea plantation worker* and *carpenter*, found in *Skilled Agricultural* and *Craft & Trades* groups, are linked through `associated_with` edges, indicating culturally familiar but normatively neutral perceptions.

5.3 ModelBiasKG Analysis

We analyze the edge statistics of each *ModelBiasKG* to characterize occupation–race bias patterns captured by SD1.5, SD2.1, SDXL, FLUX, and DALL·E 3.

Common Patterns Across Models Across all models, *White* exhibits the highest connectivity (>27.5%), followed by *Hispanic*, *Black*, and *East Asian*. Low-salience groups (*MENA*, *Indigenous*, *Mixed*, *Southeast Asian*) remain below 2.5% of all edges, indicating marginal representational presence. Most non-*White* groups are primarily connected via `underrepresented_in (R → O)` edges, suggesting systematic exclusion from occupational depictions. Stereotypes are consistently concentrated in skilled and managerial occupations, while `dominant_in` relations are virtually absent. Bias clusters further in *Clerical Support*, *Service and Sales*, and *Craft Trades*.

Per-Model Characteristics

- **SD1.5** – Stereotyped edges are rare (White: 1.5%, Black: 0.5%), but `underrepresented_in (R → O)` dominates (47%).
- **SD2.1** – *White* stereotypes dominate (6%). `Underrepresented_in (O → R)` is concentrated in *Managers* (2.5%), with no other strong occupational bias clusters.
- **SDXL** – *White* bias is expressed via `associated_with` and `stereotyped_as` edges (17.5%). *Managers* also have the highest `underrepresented_in (O → R)` frequency (2.5%).
- **FLUX** – Bias is dominated by `underrepresented_in (R → O)` (70%), with `underrepresented_in (O→R)` showing an above-average frequency only for *Managers* (2.5%).
- **DALL·E 3** – *White* dominates in both `associated_with` and `stereotyped_as` edges (17%). `Associated_with` peaks in *Skilled Agricultural* (4%), and `underrepresented_in (O → R)` in *Managers* (3.5%).

Overall, high-visibility racial groups dominate representational space across models, but the occupational bias clusters differ by architecture. This divergence—combined with consistent gaps for low-salience racial groups—highlights how model design and training data jointly shape representational bias, and where they fail to reflect culturally recognized patterns captured in SocialBiasKG.

5.4 Cross-Model Bias Analysis

While Section 5.3 analyzed models individually, here we take a comparative view, using *ModelBiasKG* to align identical *(race, occupation, directed edge)* triples across models. This enables detection of consistently reproduced biases, model-specific relations, and variations in bias intensity—revealing divergences and commonalities in representational patterns not evident from single-model analysis.

Among the triples present in all models, 94% correspond to `underrepresented_in (R → O)`, 6% to `underrepresented_in (O → R)`, 4% to `stereotyped_as (O → R)`, and 1% to `associated_with (O → R)`. By occupation group, *Manager* holds the largest share (12.4%), followed by *Technicians & Associate Professionals* and *Service & Sales Workers* (10.5% each), with the remaining groups at approximately 9.5%. This distribution suggests that, across all models, the most consistent bias tendency is the systematic underrepresentation of certain races across occupational categories (`underrepresented_in, R → O`). The relatively small proportion of `stereotyped_as` and `associated_with` edges indicates that while explicit racial stereotyping is present, it is less universally reproduced than representational gaps. The near-uniform spread across occupation groups further implies that this underrepresentation bias is not confined to specific job sectors but is pervasive across both high- and low-prestige roles.

Among the three SD models(SD1.5, SD2.1, and SDXL), no unique edges were found for SD 1.5 or SD 2.1, suggesting their inferred occupation–race biases substantially overlap with those of other models. In contrast, SDXL produced 10 unique edges, mostly `associated_with` (O $\rightarrow$ R) relations in normatively neutral or culturally specific contexts—for example, agricultural roles like *Tea plantation worker–South Asian* or service roles like *Tourist guide–Hispanic*. This indicates that SDXL tends to generate additional culturally linked associations absent in other models, potentially reflecting broader coverage of benign occupational stereotypes rather than overtly negative bias.

Across the 105 shared edges, FLUX most frequently ranked highest in bias intensity (over 80% of cases), suggesting a consistent tendency to reproduce or amplify occupation–race biases more strongly than other models. DALL·E 3 and SDXL typically occupied intermediate ranks, while SD 2.1 and SD 1.5 showed comparatively lower intensities. These results indicate that differences in training data and architecture substantially affect bias strength, and that newer or higher-capacity models do not necessarily exhibit reduced bias.

5.5 SocialBiasKG vs. ModelBiasKG Comparative Analysis

We compare each *ModelBiasKG* with the human-annotated *SocialBiasKG* to identify **SocialBiasKG-only edges**—bias links present in human perception but absent in model outputs. These gaps indicate instances where models overlook culturally salient occupation–race associations, underscoring underrepresentation or omission when compared to real-world bias patterns. Analyzing these gaps is essential for understanding not only where models fail to reflect real-world social perceptions, but also which cultural biases are systematically excluded—insights that are critical for developing fairer and more socially aware generative systems.

Many socially perceived biases—particularly `underrepresented_in (R → O)` edges for less represented races—are absent in all models. Notably, 100% of such edges for *Indigenous*, *MENA*, and *Mixed* identities in *SocialBiasKG* are missing, as are 78% for *Southeast Asian* and 65% for *South Asian*. Even for more visible groups, substantial gaps remain: 52% for *Black*, 49% for *Hispanic*, and 44% for *East Asian*. These missing biases often involve high-status professional and managerial roles where human annotators perceive systemic underrepresentation, but models fail to reflect these patterns. This absence may stem from limited training exposure to visual depictions of *MENA* and *Indigenous* individuals in diverse occupational contexts, combined with weaker cultural salience in globally aggregated datasets. As a result, models may underrepresent these identities not because the bias is absent, but because the training data fails to encode the social perception of their occupational underrepresentation. In other cases, socially recognized `associated_with` or `stereotyped_as` links (e.g., between specific minority groups and certain cultural or service-sector occupations) are entirely absent, suggesting that T2I models undercapture culturally salient, minority-specific biases while reproducing more visible, majority-centric patterns.

6 Discussion

Limitations This study focuses exclusively on evaluating occupation–race biases as represented in the human-annotated *SocialBiasKG*. While this enables controlled and culturally grounded evaluations, it limits the scope to biases already observed in social perception. Model-specific biases—those not captured in *SocialBiasKG*—may also exist and warrant independent discovery. We plan to investigate such emergent or unannotated biases in future work.

Expandability of SocialBiasKG The proposed knowledge graph framework is inherently extensible to other social bias axes, such as gender, age, religion, or socioeconomic status. While our current *SocialBiasKG* is constructed from a global, multicultural perspective, future iterations could incorporate localized knowledge graphs tailored to specific cultural contexts. This would enable comparative analyses across global and regional perspectives, highlighting how social perceptions of identity and occupation vary across societies.

7 Conclusion

We propose a socially grounded knowledge graph-based framework for evaluating occupation–race bias in Text-to-Image (T2I) models. Recognizing the limitations of purely statistical approaches, we introduce **SocialBiasKG**, a structured knowledge graph that captures nuanced, culturally informed relationships between occupations and racial groups. By modeling social perception through directed edge types—capturing various forms of social bias—SocialBiasKG enables interpretable and fine-grained bias evaluation. Building on this schema, we develop a comprehensive bias evaluation dataset and protocol that supports controlled and context-sensitive testing of T2I model behavior. Our protocol tailors evaluation strategies to the semantics of each edge type, enabling socially meaningful assessments of model outputs. Furthermore, we introduce **ModelBiasKG**, a model-derived knowledge graph that structurally summarizes representational biases in generated images, allowing direct comparison across models and against human-annotated social knowledge.

Our experiments reveal that underrepresentation biases in *SocialBiasKG*—especially for *MENA*, *Indigenous*, and *Mixed* identities—are absent across all models, exposing a systemic blind spot rooted in limited training exposure and data imbalance. Representational gaps persist more consistently than explicit stereotypes, with bias intensity varying by architecture. By detecting these culturally salient omissions and mapping where model behavior diverges from human perception, our KG-based approach advances fairness, transparency, and accountability in generative vision systems—core pillars of Trustworthy AI. Beyond occupation and race, the knowledge graph approach can be extended to other demographic attributes and social dimensions, opening new avenues for culturally responsible evaluation of multimodal AI models.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bansal, H.; Yin, D.; Monajatipoor, M.; and Chang, K.-W. 2022. How well can text-to-image generative models understand ethical natural language interventions? In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 1358–1370. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, 1493–1504.

Chakraborty, R.; Wang, Y. O.; Gao, J.; Zheng, R.; Zhang, C.; and De la Torre, F. D. 2024. Visual data diagnosis and debiasing with concept graphs. *Advances in Neural Information Processing Systems* 37:106383–106410.

Cheong, M.; Abedin, E.; Ferreira, M.; Reimann, R.; Chalson, S.; Robinson, P.; Byrne, J.; Ruppanner, L.; Alfano, M.; and Klein, C. 2024. Investigating gender and racial biases in dall-e mini images. *ACM J. Responsib. Comput.* 1(2).

Chinchure, A.; Shukla, P.; Bhatt, G.; Salij, K.; Hosanagar, K.; Sigal, L.; and Turk, M. 2024. Tibet: Identifying and evaluating biases in text-to-image generative models. In *European Conference on Computer Vision*, 429–446. Springer.

Cho, J.; Zala, A.; and Bansal, M. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3043–3054.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* 45(9):10850–10869.

Deshpande, A.; Ruiter, D.; Mosbach, M.; and Klakow, D. 2022. StereoKG: Data-driven knowledge graph construction for cultural knowledge and stereotypes. In Narang, K.; Mostafazadeh Davani, A.; Mathias, L.; Vidgen, B.; and Talat, Z., eds., *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, 67–78. Seattle, Washington (Hybrid): Association for Computational Linguistics.

D'Incà, M.; Peruzzo, E.; Mancini, M.; Xu, D.; Goel, V.; Xu, X.; Wang, Z.; Shi, H.; and Sebe, N. 2024. Openbias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12225–12235.

Franklin, J. S.; Bhanot, K.; Ghalwash, M.; Bennett, K. P.; McCusker, J.; and McGuinness, D. L. 2022. An ontology for fairness metrics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 265–275.

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

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33:6840–6851.

International Labour Office. 2012. *International Standard Classification of Occupations 2008 (ISCO-08): Structure, Group Definitions and Correspondence Tables*. Geneva: International Labour Office.

Jha, A.; Prabhakaran, V.; Denton, R.; Laszlo, S.; Dave, S.; Qadri, R.; Reddy, C.; and Dev, S. 2024. ViSAGe: A global-scale analysis of visual stereotypes in text-to-image generation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12333–12347. Bangkok, Thailand: Association for Computational Linguistics.

Ko, H.-K.; Park, G.; Jeon, H.; Jo, J.; Kim, J.; and Seo, J. 2023. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th international conference on intelligent user interfaces*, 919–933.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; Lacey, K.; Levi, Y.; Li, C.; Lorenz, D.; Müller, J.; Podell, D.; Rombach, R.; Saini, H.; Sauer, A.; and Smith, L. 2025. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space.

Li, T.; Khashabi, D.; Khot, T.; Sabharwal, A.; and Srikumar, V. 2020. UNQOVERing stereotyping biases via underspecified questions. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3475–3489. Online: Association for Computational Linguistics.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 12888–12900. PMLR.

Luo, C. F.; Ghawanmeh, A.; Zhu, X.; and Khattak, F. K. 2024a. Biaskg: Adversarial knowledge graphs to induce bias in large language models. *CoRR*.

Luo, H.; Huang, H.; Deng, Z.; Liu, X.; Chen, R.; and Liu, Z. 2024b. Bigbench: A unified benchmark for social bias in text-to-image generative models based on multi-modal llm. *arXiv e-prints* arXiv–2407.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters* 20(3):209–212.

Morales, S.; Clarisó, R.; and Cabot, J. 2025. Imagebite: A framework for evaluating representational harms in text-to-image models. In *2025 IEEE/ACM 4th International Conference on AI Engineering–Software Engineering for AI (CAIN)*, 95–106. IEEE.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. 2022. BBQ: A hand-built bias benchmark for question answering. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 2086–2105. Dublin, Ireland: Association for Computational Linguistics.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1(2):3.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35:36479–36494.

UNESCO Information for All Programme (IFAP). 2024. Mainstreaming accessibility and inclusivity in ai and digital technologies. *UNESCO*. Article describing UNESCO-IFAP's work on ensuring accessibility and inclusivity in AI technologies.

United Nations General Assembly. 2024. Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development (resolution a/78/l.49). United Nations General Assembly resolution. First ever UNGA resolution on AI; consensus adoption on 21 March 2024; defines principles for "safe, secure and trustworthy" AI.

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

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13(4):600–612.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; Hutchinson, B.; Han, W.; Parekh, Z.; Li, X.; Zhang, H.; Baldridge, J.; and Wu, Y. 2022. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.* 2022.