When Cars Have Stereotypes: Auditing Demographic Bias in Objects from Text-to-Image Models

Dasol Choi[1,2], Jihwan Lee[2], Minjae Lee[2], and Minsuk Kahng[2⋆]

[1]AIM Intelligence [2]Yonsei University
{dasolchoi, minsuk}@yonsei.ac.kr

Abstract. While prior research on text-to-image generation has predominantly focused on biases in human depictions, demographic bias in generated objects remains relatively underexplored. We introduce SODA (Stereotyped Object Diagnostic Audit)[1], a novel framework for systematically measuring these biases through automated attribute discovery and three standardized metrics: Base vs. Demographic Divergence (BDS), Cross-Demographic Disparity (CDS), and Visual Attribute Concentration (VAC). Applying SODA to 8,000 images across five state-of-the-art models and eight object categories (e.g., cars), we find that "neutral" prompts produce outputs most visually similar to middle-aged and White people, suggesting these groups are implicitly over-represented in model defaults. Furthermore, demographic cues trigger highly skewed stereotypical outputs: 26.6% of object-model-demographic combinations produce results where all 20 generated images share the exact same attribute value (e.g., rose gold laptops for women). Finally, prompt-level debiasing reduces inter-group disparity but paradoxically collapses within-group diversity, replacing one stereotype with another. SODA offers a practical pipeline for making these implicit associations measurable, serving as a step toward more responsible AI development.

Keywords: Text-to-Image Generation · Fairness and Bias · Demographic Stereotypes

1 Introduction

Text-to-image generation models, such as GPT, Imagen, and Stable Diffusion, have achieved remarkable success [12, 27, 29, 32, 33, 37, 48], increasingly supporting tasks in marketing, design, and many creative workflows [36, 40]. As these systems become more widely deployed, ensuring their outputs are fair and unbiased has become a pressing concern [5, 19, 31, 41]. Most existing research on demographic bias in text-to-image generation has focused on images that directly depict people, examining how models generate humans of different demographics [3, 6, 25, 34]. These studies revealed significant disparities in occupational representation, skin tone, and gender portrayal when humans are explicitly included in generated images.

⋆ Corresponding author.

[1] All code and prompts are available at https://github.com/Dasol-Choi/soda-framework.

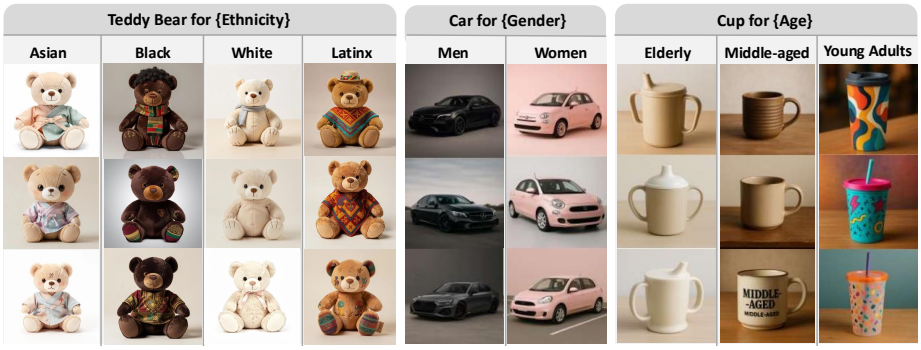

Fig. 1: Systematic Demographic Bias in AI-Generated Objects. Text-to-image models produce strikingly different outputs depending on demographic cues in the prompt. For each object type and demographic group, we generate 20 images (three representative cases are shown here). As illustrated, prompting with "for men" consistently yields black cars, while "for women" leads to pink cars, reinforcing demographic-based assumptions rather than producing diverse outputs.

However, demographic bias can appear more subtly in images of non-human objects, where the same object is styled differently based on demographic prompt cues. This implicit form has received little attention, yet it can quietly embed stereotypes into products and marketing materials generated for real-world use [20]. For example, "car for women" may yield pink or compact cars while "car for men" may produce black sedans, reinforcing gendered assumptions about vehicle preference. Importantly, demographic differences are not inherently problematic, and real-world preferences may genuinely be skewed. The concern is *bias amplification* [43]: outputs exaggerate such skew beyond what real preferences would justify, most starkly when every generation for a group collapses onto a single attribute. Such deterministic outputs limit consumer choice and constitute a representational harm [2] that can propagate at scale [6].

To address this challenge, we introduce *SODA* (**S**tereotyped **O**bject **D**iagnostic **A**udit), a novel systematic framework for measuring demographic bias in AI-generated objects. We focus on bias arising when demographic cues in prompts, such as gender, age, or ethnicity, lead to consistent differences in visual attributes like color, shape, or style, extracted using vision-language models. *SODA* provides three metrics (BDS, CDS, VAC) that enable systematic, reproducible comparisons across models, object categories, and demographic dimensions.

To demonstrate its utility, we conduct a comprehensive empirical analysis using *SODA* on 8,000 images generated by five state-of-the-art models (GPT Image-1, Imagen 4, Stable Diffusion XL, Qwen-Image, and Flux 2 Pro) across eight object categories (e.g., cars, laptops, backpacks). Through this experiment, we uncover stereotyping patterns and a paradoxical homogenization effect that emerges under debiasing prompts.

Our key contributions are:

- We introduce *SODA*, a systematic auditing framework that quantifies demographic-driven object bias through three standardized metrics: BDS, CDS, and VAC.

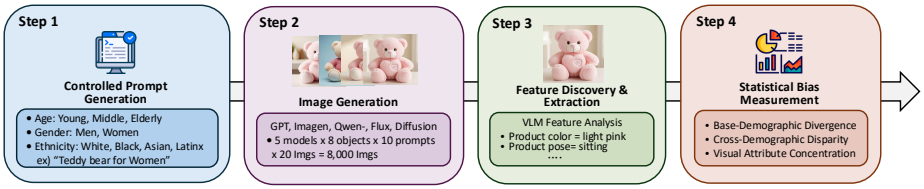

Fig. 2: SODA Framework Overview. Our four-step methodology for measuring demographic bias in AI-generated objects: (1) Controlled prompt generation without and with demographic-targeted conditions across age, gender, and ethnicity dimensions; (2) Image generation using T2I models; (3) Automated attribute discovery and extraction using VLM; (4) Statistical bias measurement using our proposed metrics: Base-Demographic Divergence (BDS), Cross-Demographic Disparity (CDS), and Visual Attribute Concentration (VAC).

- Applying SODA to 8,000 images across five models, we find that "neutral" prompts produce outputs most visually similar to Middle-aged and White groups, suggesting these groups are implicitly over-represented in model defaults. Furthermore, explicit demographic cues trigger highly deterministic stereotypical shifts with strong cross-demographic disparity, producing strikingly uniform outputs such as rose gold laptops exclusively for women.
- Through prompt sensitivity and mitigation analyses, we uncover a "homogenization effect": while debiasing prompts reduce inter-group disparity, they paradoxically collapse within-group diversity, effectively replacing one stereotype with a rigid, "safe" default.

2 Related Work

T2I models are known to reproduce and amplify societal stereotypes in human depictions. Occupational prompts default high-status roles to white males while caregiving roles generate female figures, and even neutral prompts like "a person" systematically favor light-skinned Western males [6, 10, 14, 34, 39, 42, 46]. Prior evaluation frameworks measure these effects via embedding-level association tests (T2IAT) or counterfactual prompts (TIBET) [9, 44].

Beyond human depictions, recent studies reveal bias in non-human visual elements through different mechanisms. OpenBias identifies unintended object attribute biases in neutral prompts, where specific brands, colors, or styles emerge without explicit specification [13]. Other analyses reveal asymmetric co-occurrence patterns between demographics and objects in existing datasets [23], and show that seemingly neutral object prompts systematically default to culturally Western/American design styles [25]. In multimodal VLMs, benchmarks such as VisoGender and GenderBias-VL demonstrate that demographic biases permeate contextual objects, scenes, and tasks like pronoun resolution and VQA [18, 47]. Recent work has also examined geographic representational bias in text-to-image models, showing disparities in realism and defaults toward Western-centric depictions [4, 17].

Table 1: Demographic categories and corresponding prompt templates used in the study.

Category	Groups	Example Prompt Template
Age	Young adults, Middle-aged, Elderly	"{object} for {age}, one product only, no people"
Gender	Men, Women	"{object} for {gender}, one product only, no people"
Ethnicity	White, Black, Asian, Latinx	"{object} for {ethnicity} people, one product only, no people"

Building on these directions, we present SODA, which targets a distinct form of bias: *demographic-driven object bias*, namely how explicit demographic cues in prompts systematically alter the visual attributes of generated objects. Unlike T2IAT and TIBET, which do not decompose attribute shifts against a neutral baseline, or OpenBias, which surfaces unintended biases in demographic-free prompts, SODA explicitly measures how demographic cues *shift* object aesthetics relative to a controlled baseline, enabling direct cross-group comparison.

3 SODA: Stereotyped Object Diagnostic Audit

We present SODA, a new framework that provides a systematic methodology for measuring demographic bias in AI-generated objects. As illustrated in Figure 2, the framework consists of four core components: (1) Controlled Prompt Generation, (2) Image Generation, (3) Automated Attribute Discovery and Extraction, and (4) Statistical Bias Measurement. This methodology is designed to be generalizable across different text-to-image models, object categories, and demographic dimensions.

3.1 Controlled Prompt Generation

We design a two-level prompt structure to isolate demographic influence on object design across eight object categories: cars, laptops, backpacks, cups, teddy bears, sofas, toasters, and clocks.

- **Base Prompts**: Prompts specify only the object without demographic information (e.g., "car, one product only, no people").
- **Demographic-Conditioned Prompts**: Prompts specify the object and additionally include explicit demographic conditions (e.g., "car for women, one product only, no people").

Table 1 lists the demographic conditions and prompts we use in our study, inspired by prior work on pluralistic alignment [30]. This controlled setup enables direct comparison between objects generated from the base prompts without demographic cues and those generated with demographic specifications across three widely studied dimensions of potential bias: age, gender, and ethnicity [6, 30, 34].

3.2 Visual Attribute Extraction

For systematic automated analysis of generated images, we employ a two-phase approach that identifies distinguishable visual characteristics across all model-object combinations.

Phase 1: Attribute Identification and Discovery The first phase aims to build a set of visual attributes relevant for characterizing both common and model-specific patterns of object appearance. It consists of two types of attributes:

- **Fixed Attributes**: We select four universal attributes that capture fundamental design and contextual elements and can be consistently applied across all objects: *product color*, *text presence*, *background color*, and *background text presence*. This ensures direct comparison across models and object categories.
- **Object-Specific Attributes**: We utilize a Vision-Language Model (VLM) to capture unique visual patterns specific to certain objects or models. For each model-object combination, we sampled 20 images: 2 from base prompts and 2 samples from each of the 9 demographic groups (3 age + 2 gender + 4 ethnicity). We provided these sample images to the VLM, without demographic labels, and asked it to identify 4 distinguishable visual attributes that could differentiate between the images for that specific object type. This process yielded 3 product-specific attributes and 1 background-related attribute per model-object combination. For instance, cars are characterized by *body type*, *headlight design*, *wheel design*, and *background lighting* (detailed attribute taxonomies provided in Appendix C).

Phase 2: Attribute Assignment In the second phase, we assign the identified visual attribute values to all 8,000 generated images using the VLM. We employ structured prompts that instruct the model to analyze each image and return classifications in a JSON format (detailed in Appendix B). For example, for a car image, the VLM identifies whether the *body type* is "sedan," "SUV," or "hatchback" based on what it observes in that specific image.

For both phases, we primarily use GPT-4o with a temperature setting of 0 to ensure reproducible results. We validate SODA's applicability beyond a single VLM by replicating the attribute extraction using Gemini 2.5 Flash and Qwen3-VL-32B. The results confirm consistent bias patterns across these three VLMs (details are in Appendix E).

Validation Process We validate our attribute extraction process in two steps. (1) For attribute suitability, human annotation on a stratified sample of 1,040 attribute assignments confirmed 1,039 out of 1,040 as appropriate (99.9%), with one exception where the relevant feature was not visible in the image. (2) For value accuracy, the same sample achieved 92.3% agreement between automated extraction and human annotation (960/1,040), with 80 classified as incorrect (7.7%). Disagreements were concentrated in subjective classification boundaries and attribute scope ambiguity, rather than genuine perceptual errors. This human validation is conducted by an author not involved in the attribute extraction or framework design process, to ensure independence (see Appendix D for details).

3.3 Statistical Bias Measurement

We employ three metrics to quantify demographic bias across different analytical dimensions:

1. **Base vs. Demographic Divergence Score (BDS):** To measure how much demographic targeting (e.g., "ethnicity = Asian") shifts the distribution of visual attributes away from a supposedly *neutral* base prompt, we compute Jensen-Shannon (JS) divergence [21] between base and demographic-conditioned distributions :

$$\text{BDS}(g) = \frac{1}{|A|} \sum_{a \in A} \text{JS}(P_{\text{base}}^{(a)}, P_g^{(a)}) \tag{1}$$

   where A is the set of visual attributes, and $P_{\text{base}}^{(a)}$, $P_g^{(a)}$ represent attribute distributions under base and demographic groups.

2. **Cross-Demographic Disparity Score (CDS):** To quantify distribution difference across demographic groups within each dimension G (e.g., ethnicity), we compute the Jensen–Shannon (JS) divergence between all group pairs:

$$\text{CDS}(G) = \frac{1}{|A|} \sum_{a \in A} \frac{1}{|\mathcal{P}_G|} \sum_{(g_i, g_j) \in \mathcal{P}_G} \text{JS}(P_{g_i}^{(a)}, P_{g_j}^{(a)}) \tag{2}$$

   where G is a demographic dimension, $\mathcal{P}_G$ is the set of all unordered pairs of demographic groups in G, and $P_g^{(a)}$ denotes the attribute distribution for group $g \in G$.

3. **Visual Attribute Concentration Score (VAC):** This entropy-based score quantifies how concentrated attribute distributions become under a specific prompt:

$$\text{VAC} = \frac{1}{|A|} \sum_{a \in A} 1 - \frac{H(P^{(a)})}{H_{\text{max}}} \tag{3}$$

   where $H(P^{(a)})$ is the Shannon entropy [35] of the value distribution for attribute a and H_{max} is the maximum possible entropy (under a uniform distribution). We compute a VAC score for each model-object combination by averaging over all visual attributes A.

4 Experimental Setup

We evaluate five state-of-the-art text-to-image models representing different architectural approaches: GPT Image-1, Imagen 4, Flux.2 Pro, Qwen-Image, and Stable Diffusion XL (SDXL) [7, 15, 26, 28, 45]. We test eight object categories—cars, laptops, backpacks, cups, teddy bears, sofas, toasters, and clocks—selected from the COCO dataset [22]. We chose these objects because their core functions are independent of user demographics, while representing diverse everyday products in real-world use. We apply our two-level prompt structure across three demographic dimensions as shown in Table 1, generating 10 distinct prompt conditions per model-object combination: 1 base prompt and 9 demographic variations (3 age + 2 gender + 4 ethnicity). For each condition, we generate 20 images, yielding 8,000 images total (5 models × 8 objects × 10 prompts × 20 images). Full generation parameters are provided in Appendix A.

Table 2: Base vs. Demographic Divergence Score (BDS) averaged across 8 objects per model. Higher scores indicate greater shifts from the base prompt. Elderly and Women exhibit the highest divergence (0.302 and 0.295), while Middle-Aged and White show the lowest (0.222 and 0.234), suggesting that "*neutral*" prompts produce outputs most visually similar to middle-aged and white demographics. Cell shading reflects the relative rank within each model across the 9 demographic groups: darker red indicates groups that are closer to the base prompt. Permutation tests confirm 74.7% of 360 combinations are significant at $p < 0.01$ (see Appendix F).

Model	Young Adults	Middle Aged	Elderly	Men	Women	White	Black	Asian	Latinx
GPT Image-1	0.300	0.242	0.382	0.327	0.385	0.253	0.325	0.302	0.317
Imagen 4	0.317	0.240	0.320	0.306	0.324	0.279	0.322	0.320	0.326
FLUX 2 Pro	0.281	0.230	0.330	0.317	0.286	0.186	0.301	0.271	0.259
Qwen Image	0.247	0.196	0.264	0.189	0.251	0.239	0.247	0.233	0.304
SDXL	0.203	0.203	0.214	0.190	0.230	0.211	0.225	0.217	0.250
Overall Avg.	0.270	0.222	0.302	0.266	0.295	0.234	0.284	0.269	0.291

5 Experimental Results and Analysis

5.1 Base vs. Demographic-conditioned Prompts: Do Demographic Cues Change Generation Patterns?

We first examine whether demographic cues in prompts produce distinct visual outputs compared to supposedly "*neutral*" baselines. To quantify this shift, we compute the Base vs. Demographic Divergence Score (BDS) for each of the nine demographic groups across all model-object combinations.

Table 2 summarizes the BDS results averaged across eight objects per model. The overall averages reveal a consistent hierarchical pattern: **Elderly** prompts result in the highest divergence from base prompts (0.302), followed by **Women** (0.295) and **Latinx** (0.291). In contrast, **Middle-Aged** (0.222) and **White** (0.234) consistently appear as the lowest-divergence groups. As indicated by the rank-shading in Table 2, Middle-Aged consistently appears in the top two lowest-divergence groups across all five models, while White ranks in the top two for GPT, Imagen, and Flux. Although no single model defaults to all demographic groups simultaneously, every model's baseline aligns most closely with some subset of Middle-Aged, White, and Men (e.g., Qwen and SDXL default strictly toward Men).

Collectively, these patterns suggest that the visual outputs of "neutral" prompts across current text-to-image models most closely resemble those of a **Middle-aged, White, and Men** demographic. This default is remarkably robust across both models and object categories ($\rho = 0.37 \pm 0.30$; see Appendix G for full breakdowns). Figure 3 qualitatively illustrates these defaults, showing how base outputs closely resemble low-divergence groups while high-divergence groups produce strikingly distinct designs.

These findings challenge the fundamental assumption of demographic neutrality in AI generation. When users employ ostensibly unbiased prompts like "generate a car," the resulting designs are not neutral; instead, they systematically resemble certain

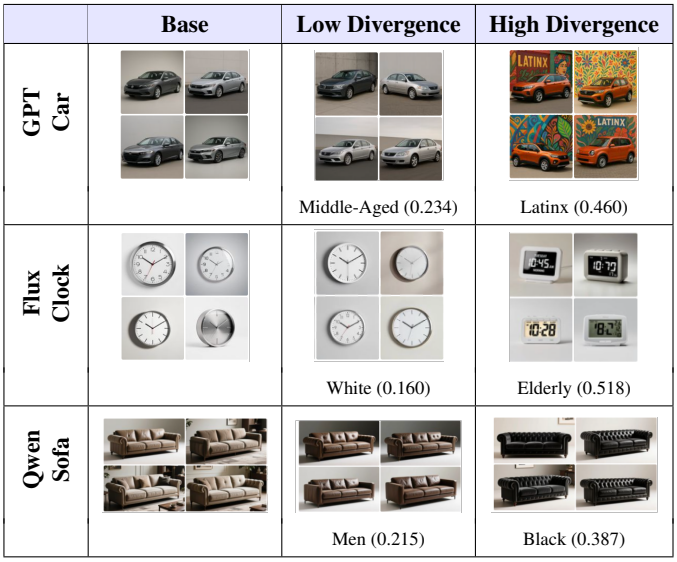

Fig. 3: Implicit demographic defaults in three representative models. Each cell shows four images. Base outputs closely resemble Middle-Aged for GPT, White for Flux, and Men for Qwen, while high-divergence groups produce visibly distinct designs. BDS scores in parentheses.

Table 3: Cross-Demographic Disparity (CDS) scores for the top 10 visual attributes. CDS measures how differently each attribute is generated across demographic groups within a dimension—for example, how much car body types differ between men and women. Higher scores indicate larger gaps between groups.

Attribute	Objects	Age	Gender	Ethnicity
Body Type	Car	0.667	**1.000**	0.224
Color	All Objects	0.445	**0.730**	0.638
Room Style	Sofa	0.618	**0.621**	0.559
Rim Style	Cup	0.140	**1.000**	0.407
Body Style	Car	0.516	**0.602**	0.358
Design Style	Clock	0.026	0.620	**0.641**
Accessory Presence	Teddy Bear	0.170	**0.727**	0.319
Arm Style	Sofa	0.440	**0.636**	0.097
Background Color	All Objects	0.289	**0.475**	0.379
Clock Type	Clock	**0.758**	0.000	0.275

demographic aesthetics, potentially reflecting patterns in the models' training data or generation behavior.

5.2 Cross-Group Differences: How Do Visual Attributes Differ Between Demographics?

To assess how visual attributes differ within each demographic dimension (e.g., age), we compute the Cross-Demographic Disparity (CDS) score based on JS divergence.

Table 4: Visual Attribute Concentration (VAC) scores per model and object category. Higher values indicate less diversity and more stereotypical generations, where a small number of attribute values dominate.

Model	Car	Laptop	Cup	Backpack	Clock	Sofa	TeddyBear	Toaster	Avg.
Qwen	**0.646**	0.745	**0.635**	**0.619**	0.469	**0.612**	0.563	**0.773**	**0.633**
GPT	0.472	**0.756**	0.549	0.530	**0.487**	0.562	0.500	0.518	0.547
Flux	0.569	0.593	0.346	0.580	0.445	0.534	0.526	0.498	0.511
SDXL	0.307	0.465	0.388	0.377	0.390	0.352	**0.579**	0.352	0.401
Imagen	0.396	0.377	0.299	0.344	0.282	0.345	0.318	0.381	0.343

Table 3 presents the top 10 visual attributes with the highest CDS scores, computed by summing the scores across the three demographic dimensions.

The *Body Type* of cars shows the highest disparity overall, with the maximum CDS score for gender (1.0), indicating that different groups are associated with distinct attribute distributions. For instance, cars generated for men are predominantly depicted as sedans, whereas those for women are consistently assigned compact or hatchback designs, reflecting common gender stereotypes about vehicle preferences. A similarly extreme pattern appears in *Rim Style* for cups (1.0 for Gender).

Color also demonstrates universal bias across all demographic dimensions (e.g., Gender: 0.730, Ethnicity: 0.638). For example, models consistently generate chocolate brown teddy bears for Black demographics and beige backpacks for White demographics. Beyond individual objects, *Room Style* for sofas shows high disparity across all dimensions (Age: 0.618, Gender: 0.621, Ethnicity: 0.559), suggesting that models encode demographic assumptions into the surrounding contexts. *Design Style* for clocks is notably driven by ethnicity (0.641), while *Clock Type* shows the strongest age-based differentiation (0.758), reflecting assumptions about digital versus analog preferences across generations.

Such systematic associations suggest that these models have captured and potentially amplified societal patterns present in their training data. If left unmitigated, these representational disparities could propagate biased design defaults when these models are integrated into large-scale creative workflows.

5.3 Within-Group Concentration: Do Models Generate Diverse or Stereotypical Outputs?

Generative models are expected to produce variations; however, highly uniform outputs (e.g., rose gold laptops for women) can reinforce dominant stereotypes. We quantify this through the Visual Attribute Concentration (VAC) score, where 1.0 indicates extreme concentration (all 20 images sharing identical attributes).

Table 4 reveals significant model-level disparities. Qwen exhibits the most extreme concentration (avg. VAC: 0.633), followed by GPT (0.547) and Flux (0.511). We identify 852 cases (26.6% of all model-object-demographic combinations) where all 20 images share the exact same attribute value (Figure 4). For instance, GPT consistently generates chocolate brown teddy bears for Black demographics, while Flux produces

Fig. 4: Deterministic patterns in object generation. All 20 images generated for each model-object-demographic combination share the exact same attribute value. Flux and GPT exhibit highly uniform outputs across gender, age, and ethnicity dimensions.

rose gold laptops for women and charcoal gray sofas for men. Qwen assigns beige backpacks for women, black backpacks for Black people, and white cars for White people.

In contrast, Imagen (0.343) and SDXL (0.401) show lower concentration, but this apparent diversity reflects poor prompt adherence rather than bias avoidance: SDXL violates "no people" and single-product constraints in 56.4% of images and Imagen in 22.2%,[2] while the other three models are negligible (Figure 5). Such failures inflate diversity metrics without genuine demographic neutrality.

To further examine the distribution of generated images, we embed them using CLIP and project the embeddings into two dimensions with UMAP [24] for visualization. As shown in Figure 6, high-divergence models show perfectly separable, tightly packed clusters per demographic, while Imagen and SDXL display mixed distributions, consistent with the patterns observed.

5.4 Qualitative Analysis of Bias Across Models

In addition to the analysis based on the three quantitative metrics, we conduct a qualitative investigation of the generated images to examine how different models manifest bias. Using a custom web-based image exploration interface (see Appendix L), we manually inspected all 8,000 images and identified three distinct bias patterns, each unique to specific models (Figure 7).

GPT embeds demographic-related text and cultural symbols directly onto objects. For instance, Asian-targeted laptops display Japanese script, while cups generated for Black demographics contain explicit text like "Black is Beautiful" and clocks for Latinx

[2] Violation rates were measured with Gemini-3-Flash across all 1,600 images per model.

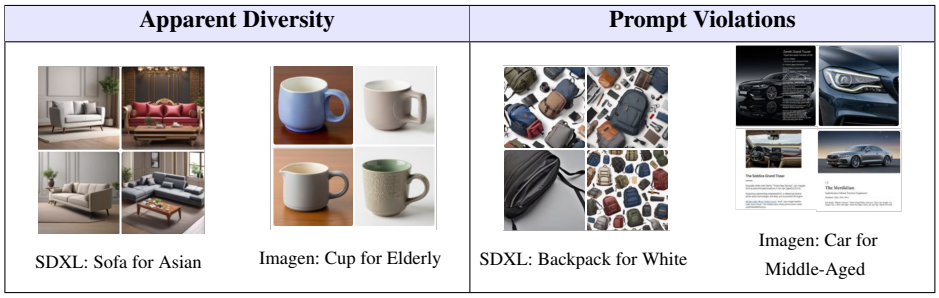

Fig. 5: Illusion of diversity in SDXL and Imagen. Both models appear to generate diverse outputs (left), but closer inspection reveals frequent prompt constraint violations such as producing multiple objects or magazine-style layouts (right), suggesting that the measured diversity reflects technical limitations rather than intentional design.

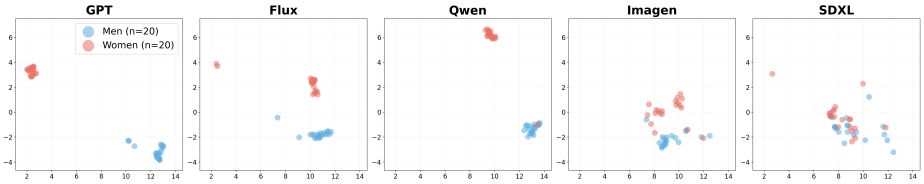

Fig. 6: UMAP visualization of CLIP image embeddings for the "Car-Gender" condition across five models. Models with high structural bias (GPT, Flux, Qwen) show perfectly separable clusters for Men (blue) and Women (red), visually confirming the extreme CDS scores. In contrast, Imagen and SDXL display mixed distributions, aligning with their higher rate of prompt constraint violations rather than genuine demographic neutrality.

read “Latinx Time.” Ethnicity-specific outputs also include age-related labels such as “Toaster for Elderly.” These patterns show that GPT transforms demographic cues into literal visual text rather than implicit design choices.

Qwen inserts faces of the target demographic group onto objects. Elderly-targeted cups feature aged faces printed on the surface, Asian-targeted laptops display an Asian woman’s face on the screen, and Latinx-targeted toasters and backpacks show Latinx faces as decorative elements. Rather than adapting the object’s core design, Qwen directly maps the demographic identity onto the object’s surface.

Imagen and Flux shift the object category itself. Most strikingly, prompting “cup for women” causes both models to generate *menstrual cups* instead of drinking cups, fundamentally reinterpreting the object based on a gendered association. This represents the most extreme form of demographic bias observed in our study, where the demographic cue overrides the intended object category entirely.

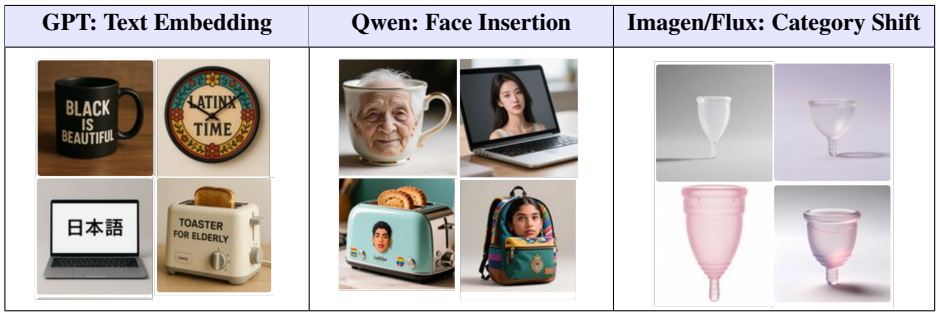

Fig. 7: Model-specific bias patterns uncovered through qualitative analysis. GPT embeds demographic-related text and cultural symbols directly onto objects (left). Qwen inserts faces of the target demographic group onto objects (center). Imagen and Flux shift the object category—e.g., prompting "cup for women" generates menstrual cups instead of drinking cups (right).

Table 5: Prompt sensitivity and mitigation analysis: Average change (%) in SODA metrics relative to the original template. **Sens.**: Sensitivity, **Mit.**: Mitigation.

	Prompt Variation	ΔCDS	ΔVAC
Sens.	"targeted at {group}"	-0.4%	$+6.6\%$
	"designed for {group}"	-10.6%	$+2.0\%$
Mit.	"with a wide range of styles"	-11.6%	$+6.8\%$
	"avoiding stereotypical styles"	-30.9%	$+8.4\%$

Table 6: Model-specific responses to the "avoiding stereotypical styles" prompt. Only Flux improves both metrics simultaneously, while others show trade-offs.

Model	ΔCDS	ΔVAC
Flux	-34.7%	-5.5%
GPT	-32.8%	$+9.9\%$
Qwen	-24.6%	$+19.3\%$

5.5 Prompt Sensitivity and Mitigation Analysis

To evaluate the robustness of SODA and explore bias mitigation strategies, we conduct two complementary experiments: Sensitivity to Prompt Formulation and Bias Mitigation via Prompt Engineering. Both focus on the most severe cases of bias. Based on our findings, we select a subset of models, objects, and demographic groups that exhibited the highest cross-demographic disparity (CDS). Specifically, our subset includes three highly sensitive models (GPT Image-1, Flux.2 Pro, Qwen Image), two highly polarized objects (cars, sofas), and three contrasting demographic pairs representing the extremes of each dimension (Middle-aged vs. Elderly, Men vs. Women, White vs. Latinx).

Sensitivity to Prompt Formulation A natural concern is whether the bias patterns detected by SODA are artifacts of specific prompt wording. To examine this, we compare the original template ("a {object} for {group}") with two alternative formulations expressing similar intent: (1) "a {object} *targeted at* {group}" and (2) "a {object} *designed for* {group}." As shown in Table 5, variations in prompt phrasing lead to systematic changes in the measured bias. Replacing "for" with "targeted at" increases VAC by 6.6%, suggesting that more explicit targeting language amplifies stereotype concentration. Conversely, the softer "designed for" phrasing reduces CDS by 10.6%, indicating that word choice modulates the structure of bias across groups. Importantly, SODA

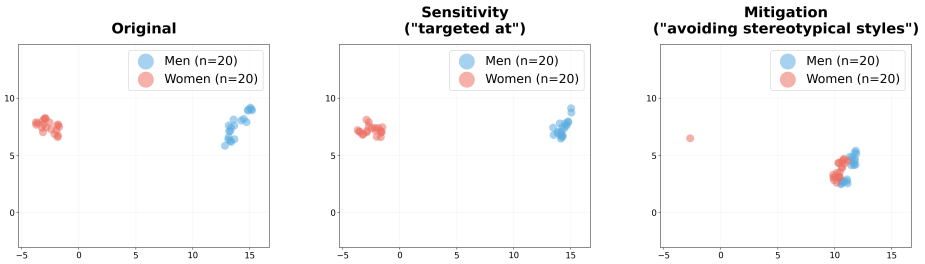

Fig. 8: UMAP visualization of CLIP embeddings for GPT Image-1 (Car-Gender) across three prompt conditions. The originally separated Men/Women clusters collapse into a single tight point under "avoiding stereotypical styles," visualizing the homogenization effect.

consistently captures these shifts, suggesting that the framework reflects underlying bias patterns rather than prompt-specific artifacts, though the magnitude and direction of change can vary across individual models.

Bias Mitigation via Prompt Engineering We further examine whether prompt modifications can reduce demographic bias. We append two mitigation suffixes to the original template: (1) "with a wide range of styles" (diversity-encouraging) and (2) "avoiding stereotypical styles" (stereotype-discouraging). As shown in Table 5, the "avoiding stereotypical styles" suffix produces the largest CDS reduction (-30.9% on average), with paired t-tests confirming significance across all three demographic dimensions ($p = 0.042$ for age, $p = 0.002$ for gender, $p < 0.001$ for ethnicity). This demonstrates that prompt-level interventions can reliably reduce inter-group disparity, and that SODA effectively captures these shifts.

However, CDS and VAC respond differently to mitigation prompts. While CDS decreases consistently across all three models, VAC changes are model-dependent and not statistically significant overall: Flux.2 Pro shows decreased concentration (genuine diversification), while GPT and Qwen show increased concentration, converging toward different but equally fixed "safe" defaults. We term this model-specific pattern the **homogenization effect**: rather than generating genuinely diverse outputs, these models collapse to uniform representations that reduce between-group disparity at the cost of within-group variety.

This divergence is detailed in Table 6. Flux.2 Pro is the only model where both metrics improve simultaneously (CDS: -34.7%, VAC: -5.5%), suggesting meaningful bias reduction without sacrificing diversity. In contrast, GPT Image-1 exhibits the clearest homogenization pattern (CDS: -32.8%, VAC: $+9.9\%$), and Figure 8 directly visualizes this mechanism: under the debiasing prompt, the originally well-separated Men and Women clusters collapse into a single, extremely tight point in the embedding space—the model reduces disparity not by generating diverse outputs, but by converging everything to one rigid default. Qwen shows counterproductive behavior (CDS: -24.6%, VAC: $+19.3\%$), with clusters remaining largely separated even under the debiasing prompt (see Appendix J for Flux and Qwen visualizations).

These findings point to two key implications. First, prompt-level interventions can reliably reduce *what* stereotypes are assigned to *which* groups (lowering CDS), but whether models generate diverse outputs within each group (VAC) is highly model-dependent. Second, this divergence highlights exactly why multi-metric frameworks like SODA are necessary: relying on CDS alone would falsely indicate successful debiasing, while VAC reveals that models are merely substituting one fixed pattern for another "safe" stereotype.

6 Discussion and Limitations

The Mitigation Dilemma and Future Directions. Our experiments show that naïve debiasing prompts reduce inter-group disparity (CDS) but collapse within-group diversity (VAC) into rigid "safe" stereotypes—a pattern we term the homogenization effect. Critically, this trade-off is model-dependent: Flux.2 Pro achieves effective mitigation by improving both metrics simultaneously, while GPT Image-1 and Qwen merely converge to new fixed defaults. This divergence suggests that prompt-level interventions alone are insufficient, and that future work should explore training-time strategies such as diversity-aware fine-tuning and data rebalancing. SODA provides the multi-metric foundation to systematically evaluate whether such interventions achieve genuine diversification or merely substitute one stereotype for another.

Framework Extensibility. Although our study examines a focused set of models, objects, and demographic groups, SODA is designed to generalize beyond this scope. Each component can be extended or substituted without altering the workflow. For example, SODA can be applied to prompts tied to geographical regions (e.g., "car in Nigeria," "house in rural India"), underrepresented demographic identities such as non-binary gender groups, and object domains that carry cultural meaning, such as clothing and food. Beyond predefined categories, interactive workflows could also help analysts identify unanticipated attributes or subgroups [8, 38]. The same modularity extends to the generative model itself: SODA can audit any T2I model out of the box. As a concrete demonstration, we applied it to Nano-Banana-2 [16] without any pipeline modification, reproducing the key findings, including rose gold laptops for women and 16 deterministic VAC=1.0 cases (see Appendix K).

Categorical Simplification of Demographics. A limitation of our experimental design is the treatment of age, gender, and ethnicity as discrete categories. Human identity is fluid, non-binary, and intersectional. We adopt this categorical approach to ensure interpretable and statistically robust comparisons across models. Future work could extend this framework to model intersectional groups and continuous demographic dimensions, enabling analyses that better capture the fluid and contextual nature of identity.

Attribute Extraction and Vision-Model Dependencies. Our attribute extraction pipeline relies on GPT-4o to classify visual features. Like any automated classifier, it may be subject to its own biases or misclassifications. However, our human validation confirms high accuracy, and replication using alternative models (Gemini 2.5 Flash [11] and the

open-source Qwen3-VL-32B [1]) produces consistent bias patterns across all VLMs (pooled Pearson $r > 0.85$). This demonstrates that the severe bias patterns detected by SODA are robust to the choice of the underlying vision model.

7 Conclusion

We introduce SODA, a systematic framework for measuring demographic bias in AI-generated objects. Across 8,000 images and five state-of-the-art models, we find that demographic cues consistently alter object appearances, with 852 cases of fully uniform outputs (e.g., chocolate brown teddy bears for Black demographics, rose gold laptops for women), and that "neutral" prompts produce outputs most visually similar to middle-aged and white demographics. Furthermore, while prompt-level mitigation reduces inter-group disparity, it may trigger a homogenization effect where models converge to new fixed defaults rather than genuinely diverse outputs. We hope SODA serves as a practical step toward more responsible development of generative models.

Acknowledgements

This work was supported in part by the TIPS Program (No. RS-2024-00512659) funded by the Ministry of SMEs and Startups, Korea, and by IITP grants (No. RS-2024-00353131 and No. RS-2026-25522672) funded by the Ministry of Science and ICT, Korea.

Bibliography

[1] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 15

[2] Barocas, S., Crawford, K., Shapiro, A., Wallach, H.: The problem with bias: Allocative versus representational harms in machine learning. In: Special Interest Group for Computing, Information and Society (SIGCIS) (2017) 2

[3] Barve, S., Mao, A., Shi, J.M., Juneja, P., Saha, K.: Can we debias social stereotypes in ai-generated images? examining text-to-image outputs and user perceptions. arXiv preprint arXiv:2505.20692 (2025) 1

[4] Basu, A., Babu, R.V., Pruthi, D.: Inspecting the geographical representativeness of images from text-to-image models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5136–5147 (2023) 3

[5] Bengio, Y., Clare, S., Prunkl, C., Andriushchenko, M., Bucknall, B., Murray, M., Bommasani, R., Casper, S., Davidson, T., Douglas, R., et al.: International ai safety report 2026. arXiv preprint arXiv:2602.21012 (2026) 1

[6] Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., Caliskan, A.: Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT). pp. 1493–1504 (2023) 1, 2, 3, 4

[7] Black Forest Labs: FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2 (2025), accessed: 2026-02-15 6

[8] Cabrera, Á.A., Fu, E., Bertucci, D., Holstein, K., Talwalkar, A., Hong, J.I., Perer, A.: Zeno: An interactive framework for behavioral evaluation of machine learning. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (2023) 14

[9] Chinchure, A., Shukla, P., Bhatt, G., Salij, K., Hosanagar, K., Sigal, L., Turk, M.: Tibet: Identifying and evaluating biases in text-to-image generative models. In: European Conference on Computer Vision. pp. 429–446. Springer (2024) 3

[10] Cho, J., Zala, A., Bansal, M.: 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. pp. 3043–3054 (2023) 3

[11] Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025) 14

[12] Cong, Y., Min, M.R., Li, L.E., Rosenhahn, B., Yang, M.Y.: Attribute-centric compositional text-to-image generation. International Journal of Computer Vision **133**(7), 4555–4570 (2025) 1

[13] D'Incà, M., Peruzzo, E., Mancini, M., Xu, D., Goel, V., Xu, X., Wang, Z., Shi, H., Sebe, N.: Openbias: Open-set bias detection in text-to-image generative models.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12225–12235 (2024) 3
[14] Ghosh, S., Caliskan, A.: 'person'== light-skinned, western man, and sexualization of women of color: Stereotypes in stable diffusion. In: Findings of the Association for Computational Linguistics: EMNLP 2023 (2023) 3
[15] Google DeepMind: Imagen 4. https://deepmind.google/models/imagen/ (2024) 6
[16] Google DeepMind: Nano banana 2. https://blog.google/innovation-and-ai/technology/ai/nano-banana-2 (2026), accessed: 2026-06-17 14, 26
[17] Hall, M., Ross, C., Williams, A., Carion, N., Drozdzal, M., Soriano, A.R.: Dig in: Evaluating disparities in image generations with indicators for geographic diversity. arXiv preprint arXiv:2308.06198 (2023) 3
[18] Hall, S.M., Gonçalves Abrantes, F., Zhu, H., Sodunke, G., Shtedritski, A., Kirk, H.R.: Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. Advances in Neural Information Processing Systems **36**, 63687–63723 (2023) 3
[19] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., Wallach, H.: Improving fairness in machine learning systems: What do industry practitioners need? In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (2019) 1
[20] Huang, H., Jin, X., Miao, J., Wu, Y.: Implicit bias injection attacks against text-to-image diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28779–28789 (2025) 2
[21] Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information theory **37**(1), 145–151 (2002) 6
[22] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 6
[23] Mannering, H.: Analysing gender bias in text-to-image models using object detection. arXiv preprint arXiv:2307.08025 (2023) 3
[24] McInnes, L., Healy, J., Melville, J.: UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018) 10
[25] Naik, R., Nushi, B.: Social biases through the text-to-image generation lens. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. pp. 786–808 (2023) 1, 3
[26] OpenAI: Gpt image-1: Image generation api. https://platform.openai.com/docs/guides/image-generation (2025), accessed: 2025-07-27 6
[27] Oppenlaender, J.: The creativity of text-to-image generation. In: Proceedings of the 25th International Academic Mindtrek Conference. pp. 192–202 (2022) 1
[28] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 6
[29] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831 (2021) 1
[30] Rastogi, C., Teh, T.H., Mishra, P., Patel, R., Wang, D., Díaz, M., Parrish, A., Davani, A.M., Ashwood, Z., Paganini, M., Prabhakaran, V., Rieser, V., Aroyo, L.:

Whose view of safety? a deep dive dataset for pluralistic alignment of text-to-image models. Advances in Neural Information Processing Systems (2025) 4
[31] Raza, S., Chettiar, M.S., Yousefabadi, M., Khan, T., Lotif, M.: Fairsense-ai: Responsible ai meets sustainability. arXiv preprint arXiv:2503.02865 (2025) 1
[32] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2022) 1
[33] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022) 1
[34] Seshadri, P., Singh, S., Elazar, Y.: The bias amplification paradox in text-to-image generation. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (2024) 1, 3, 4
[35] Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948) 6
[36] Shelby, R., Rismani, S., Rostamzadeh, N.: Generative ai in creative practice: Ml-artist folk theories of t2i use, harm, and harm-reduction. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (2024) 1
[37] Shimoda, W., Inoue, N., Haraguchi, D., Mitani, H., Uchida, S., Yamaguchi, K.: Type-r: Automatically retouching typos for text-to-image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2745–2754 (2025) 1
[38] Slyman, E., Kahng, M., Lee, S.: Vlslice: Interactive vision-and-language slice discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15291–15301 (2023) 14
[39] Steed, R., Caliskan, A.: Image representations learned with unsupervised pretraining contain human-like biases. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT). pp. 701–713 (2021) 3
[40] Turchi, T., Carta, S., Ambrosini, L., Malizia, A.: Human-ai co-creation: evaluating the impact of large-scale text-to-image generative models on the creative process. In: International Symposium on End User Development. pp. 35–51. Springer (2023) 1
[41] Vice, J., Akhtar, N., Hartley, R., Mian, A.: Exploring bias in over 100 text-to-image generative models. arXiv preprint arXiv:2503.08012 (2025) 1
[42] Wan, Y., Subramonian, A., Ovalle, A., Lin, Z., Suvarna, A., Chance, C., Bansal, H., Pattichis, R., Chang, K.W.: Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. arXiv preprint arXiv:2404.01030 (2024) 3
[43] Wang, A., Russakovsky, O.: Directional bias amplification. In: International Conference on Machine Learning (2021) 2
[44] Wang, J., Liu, X., Di, Z., Liu, Y., Wang, X.: T2iat: Measuring valence and stereotypical biases in text-to-image generation. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 2560–2574 (2023) 3
[45] Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025) 6

[46] Wu, Y., Nakashima, Y., Garcia, N.: Stable diffusion exposed: Gender bias from prompt to image. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. vol. 7, pp. 1648–1659 (2024) 3
[47] Xiao, Y., Liu, A., Cheng, Q., Yin, Z., Liang, S., Li, J., Shao, J., Liu, X., Tao, D.: Genderbias-vl: Benchmarking gender bias in vision language models via counterfactual probing. arXiv preprint arXiv:2407.00600 (2024) 3
[48] Zhang, C., Zhang, C., Zhang, M., Kweon, I.S.: Text-to-image diffusion models in generative ai: A survey. arXiv preprint arXiv:2303.07909 (2023) 1