

On the Representation Bias of Disabilities in Generative Text-to-Image Models

Asfandiyar Azhar^{*1,5}, Zaid Khan^{*2}, Shaurjya Mandal³, Nidhish Shah⁴,
Alaa Youssef⁵, Oluwasanmi Adenaiye⁶, Curtis Langlotz⁵, Brad E. Dicianno⁶

¹Department of Biomedical Engineering, Carnegie Mellon University, Pennsylvania, USA,

²School of Medicine, Eastern Virginia Medical School, Virginia, USA,

³Mass General Hospital, Harvard Medical School, Massachusetts, USA,

⁴Department of Computer Science, Eindhoven University of Technology, Eindhoven, Netherlands,

⁵Department of Radiology, Stanford University School of Medicine, California, USA,

⁶Department of Physical Medicine & Rehabilitation, University of Pittsburgh School of Medicine, Pennsylvania, USA

^{*}Equal contribution.

Correspondence: dicianno@pitt.edu

Abstract

Emerging text-to-image (T2I) models perpetuate biases, creating stereotypical portrayals of disability. This study investigates whether an agent-based, human-in-loop, AI workflow, that iteratively refines user-provided queries, can produce fairer representations of disability compared to standard models like Stable Diffusion, DALL·E-3, and MidJourney. Sixteen raters (eight with disabilities and eight without) evaluated images generated from models, prompted with textual inputs that targeted professional, medical, or social contexts. In an initial survey, raters were asked if an image depicted a disability, the type of disability if present, and if the setting aligned with the given text prompt to the model. Qualitative comments were collected. A post-survey asked raters to choose which image of a pair (standard vs. agent-based workflow) more accurately represented disability. Results revealed that standard models disproportionately produced images featuring mobility impairments or wheelchairs ($p < 0.001$), consistent with broader dataset biases. In contrast, the agent-based model's images were deemed more representative in approximately two-thirds of comparisons ($p < 0.0001$), especially among raters with disabilities. Qualitative analyses underscored the need for nuanced, contextually rich portrayals. These findings suggest disability bias and that targeted debiasing can help mitigate stereotypes in generative AI, fostering more equitable outputs.

1 Introduction

As Artificial Intelligence (AI) permeates everyday life, questions arise about its effects on marginalized communities. Existing work has shown that AI systems often inherit and perpetuate the implicit biases present in their training data (Bianchi et al.,

2023), (Caliskan et al., 2017), (Ghosh and Caliskan, 2023), (Orgad et al., 2023), (Petroni et al., 2019). Disability is frequently misrepresented in large-scale datasets, leading to marginalization of people with disabilities in AI training (Guo et al., 2019), (Mack et al., 2024), (Newman-Griffis et al., 2024). Data curation itself is not neutral; selective or incomplete representation exacerbates model biases (Guo et al., 2019), (Mack et al., 2024). For instance, LAION-5B, a 5-billion image-caption AI training dataset, has unknown completeness regarding disability (Schuhmann et al., 2022). Separately, natural language processing (NLP) models have been shown to contain disability-related biases (Hutchinson et al., 2020a). Such compounding factors may lead AI systems to underrepresent or stereotype disability in images for patient and medical education or research.

Prior AI debiasing research primarily focused on race or gender, not disability (Ghosh and Caliskan, 2023), (Schramowski et al., 2022). In Text-to-Image (T2I) generation, bias mitigation efforts have centered on improved detection metrics (Hutchinson et al., 2020b), (Lee et al., 2023) and developing fairer model architectures and training. Examples include prompt-engineering strategies that promote equitable outputs (Bianchi et al., 2023), (Orgad et al., 2023) and models like Fair Diffusion that modify diffusion to reduce bias (Friedrich et al., 2023). Other work, such as Park et al. (Park et al., 2024), has sought to deepen interpretability through visual analyses of the denoising process at each diffusion time step, potentially allowing for bias detection.

Previous works address facets of bias but not how or why disability stereotypes emerge and persist in generated images. Further, engaging with

individuals with disabilities in both assessment and design of models remains critical. Previous studies involving such individuals have presented varied perspectives on how disabilities should (or should not) be depicted in AI, particularly in the case of “less visible” disabilities (Guo et al., 2019), (Mack et al., 2024), (Newman-Griffis et al., 2024). These findings support performing inclusive evaluations, incorporating both quantitative and qualitative methods.

Accordingly, we introduce an agent-based AI workflow that refines disability related textual-inputs for debiasing. We adopt an approach inspired by previous research advocating inclusive user feedback (agent-based) (Guo et al., 2019), (Mack et al., 2024), (Newman-Griffis et al., 2024), combining systematic metrics to gauge bias in generations with analyses informed by participants with and without disabilities. This paper investigates disability bias in commonly used T2I models and evaluates the effectiveness of our workflows debiasing. Our aims and hypotheses (H) were:

Aim 1: Determine the extent to which the models represent “people with disabilities” as “people.”

- H1a: No model will generate images portraying people with disabilities when provided generic prompts.
- H1b: Human raters will identify disabilities at a higher rate when the prompt explicitly includes the word ‘disability’, and this identification rate will be highest among raters with a disability.

Aim 2: Determine the breadth of disability types and assistive technologies portrayed by mainstream T2I models.

- H2a: Physical disability will be most commonly depicted compared to other types of disabilities.
- H2b: Wheelchairs will be the most frequently depicted assistive device depicted (i.e., show “disability = wheelchair” bias).

Aim 3: Evaluate whether T2I models under-represent people with disabilities in active or provider roles across professional, medical, and social settings.

- H3: Relative to those without disabilities, images of people with disabilities will less frequently depict them in provider roles.

Aim 4: Assess whether an innovative prompt-clarification and refinement workflow reduces representational bias and yields images preferred by human raters.

- H4a: Relative to standard models, a refined AI workflow mitigates “disability = wheelchair” bias per KL-Divergence.
- H4b: In head-to-head comparisons, a majority of the 16 human raters will select images generated by the agent-based workflow over the standard model images.

2 Methods & Experiments

This study was excluded from University of Pittsburgh Institutional Review Board Review because it did not meet the definition of human subjects research. The research team consisted of university affiliated individuals with varying levels of ability. The study parameters and constructed elements such as rule-based rubric were based on previous works cited throughout the manuscript and the foundational ADA framework.

2.1 Diffusion and Debiasing Model

Diffusion models generate images by gradually converting an image x_0 to noise (forward diffusion) and then re-synthesizing an image using a learned, reverse process (p_θ). This denoising can be expressed as $p_\theta(x_{t-1}|x_t)$, where a noisy image (x_t) is incrementally de-noised to a new x_0 . Each de-noising step can be conditioned with textual embeddings (c), such as a user provided prompts for T2I models, expressed as $p_\theta(x_{t-1}|x_t, c)$. As a result, the final image reflects the model’s training distribution and provided text.

The expressive power of diffusion models often remains untapped due to poor prompting, especially regarding the specifics of disability (e.g., type, visibility, assistive devices). This results in models relying on and further entrenching stereotypes. To address this, our AI workflow strategy (Figure 1) systematically clarified prompts with a human in the loop (HITL), i.e. ‘agent-based’, before beginning image generation.

Our agent’s critic refined initial prompt c_0 into a comprehensive c^* through targeted user queries, altering the reverse process $p_\theta(x_0|c^*)$. To detect ‘bias’ for refinement, the critic employed two methods: 1) an explicit, rule-based rubric was used to flag stereotypes and keywords, and 2) the language

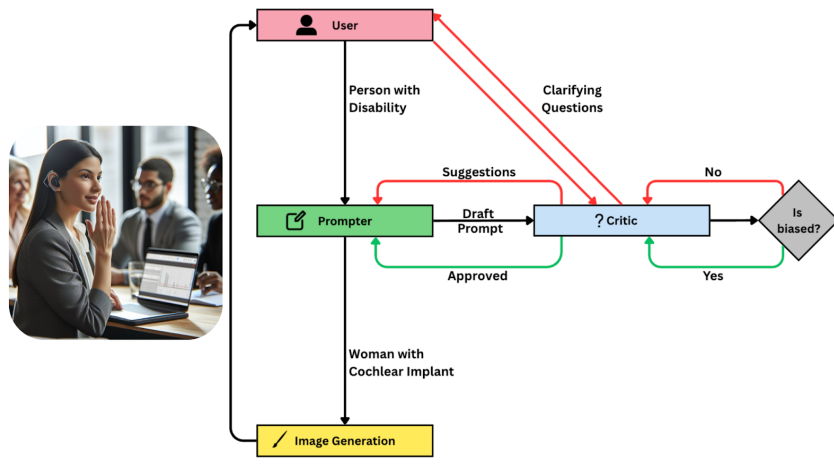


Figure 1: Agent-based Workflow for Debiasing. A human-in-the-loop framework for generating unbiased images. Users provide open-ended prompts, which are refined through an iterative process involving a critic for bias detection, and an image generation model (painter). The critic ensures fairness by flagging biases, and iterative user feedback helps improve prompt specificity. The final output is a debiased image that aligns with user intent while promoting inclusivity.

model’s existing pattern recognition was leveraged to detect negative disability associations not defined in the rubric. Example workflow in Appendix A.

This workflow was selectively applied to DALL-E 3, Stable Diffusion, and Midjourney. For prompt standardization, only disability variants from our initial prompt testing phase were used. (X only = person with disability), with two new additional prompts:

“Generate a photo of a person with disability” and “Generate a photo of a person without disability”

to address less descriptive cases. Individuals interacted with the workflow only as necessary i.e., if the agent-based system requested clarification.

We generated 60 unique images (30 standard, 30 debiased) for each of the 8 matched pairs of raters to prevent rater fatigue, totaling 480 images

We assessed how the workflow’s clarifications altered the model output by comparing $p_{\theta}(x_0|c)$ and $p_{\theta}(x_0|c^*)$ of the final images through Kullback-Leibler divergence (KLD):

$$D_{KL}(p_{\theta}(x_0|c^*) || p_{\theta}(x_0|c)) \quad (1)$$

A larger value indicates clarifying details in c^* systematically alter generated image distributions. This assessment strategy evaluated

H4a by a one-sample t-test of

$$H_0 : \mu_{KL} = 0, H_A : \mu_{KL} > 0 \quad (2)$$

where μ_{KL} is mean KL-Divergence across contexts.

2.2 Prompts & Image Generation

Four, paid subscription, diffusion-based models, DALL-E 3, Stable Diffusion, Fair Diffusion, and Midjourney, were used to generate a broad sample of images. Standardized prompts were designed to encompass each model’s handling of disability representation in realistic contexts. Prompts were:

"Generate a photo of a X in a Y setting,"

- Where X was either "person" or "person with a disability."
- Where Y was "professional," "medical," or "social."

This 6-prompt set compared model responses to ‘generic’ versus explicit disability queries across three environments. Each model used the same prompt set, enabling evaluation of how effectively they captured disability in their generations. Submitting all 6 prompts into each model created 24 images. Repeating this 8 times generated 192 images for evaluation.

2.3 Rating T2I Generations

Presurvey. We recruited 16 participants (8 matched pairs), pairing individuals with and without disabilities. Rater pairs evaluated the same



Figure 2: Visual differences across T2I model generations. Randomly sampled images generated by each model per situational context.

set of images independently and were compared with one another afterwards. Participants were academic or adjacent staff working in a university setting. Though informed the study involved images concerning disabilities, they were we kept them unaware of the specific models used and the study hypotheses specifics (i.e., 'blinded' in ableist language). Participants in each dyad received the same 24 images. For this study, disability was defined for raters as “a physical or mental impairment substantially limiting major life activities, a history of such impairment (e.g., cancer in remission), or being perceived by others as having such impairment (e.g., severe scars)” ([Americans with Disabilities Act, 1990](#)) and that “disability may not be visible and could be implied.” In addition to the human raters, GPT-4 Vision ([OpenAI, 2023](#)) was also provided with all images and questions, functioning as a separate rater.

Data was gathered via an online survey, which presented each participant with the images and questions to rate. The complete list of questions for images is in [Appendix B](#). All responses were stored in a confidential file, which recorded each de-identified participant’s disability status alongside their ratings. Human raters were queried about the presence or absence of disability (H1a, H1b), the type of disability present (H2a), relevant assistive devices (H2b), and the setting or roles being depicted in each image (H3). Last, an open-ended question was included allowing for comment on themes noted, later used for qualitative analysis.

Descriptive statistics were used to address H1a, measuring the number of images rated as display-

ing disabilities per group. Cohen’s Kappa (κ) tested H1b, measuring prompt vs rater agreement. χ^2 Goodness of Fit tests addressed H2a and H2b, determining the most common disability type and assistive devices depicted, respectively. For H3, χ^2 and Fisher’s Exact Tests were used to determine whether people with disabilities in images were more likely to be depicted as a provider rather than as a recipient of services.

Post-Survey (After Debiasing). As mentioned, raters were presented with 8 pairs of 60 images for head-to-head comparison and asked to choose the image they felt best represented a person with a disability (H4). Raters answered questions on image bias factors ([Lee et al., 2023](#)), as shown below, uninformed of the specific AI generation prompt or specific model used:

“Which image, in your opinion, better represents a person with a disability? Some factors to consider are image quality, bias, fairness, diversity, and inclusivity. Is the scenario of the person with a disability realistic? Is it clear who the person with a disability is within the image? Does the better image align with your personal definition of disability?”

Following image preference selection and completing related questions, open-ended feedback was collected for qualitative analysis.

H4a was tested using a binomial test to determine overall image preference (biased vs. debiased). For H4b, McNemar’s tests compared

Prompt	Rater Type	No Disability Shown N (%)	Disability Shown N (%)
Generic	No Disability	70 (72.9)	26 (27.1)
	With Disability	73 (76.0)	23 (24.0)
	GPT-4	90 (93.8)	6 (6.2)
Disability	No Disability	5 (5.2)	91 (94.8)
	With Disability	2 (2.1)	94 (97.9)
	GPT-4	2 (2.1)	94 (97.9)

Table 1: Comparison of rater groups on whether images showed disability. ‘Generic’ prompts did not mention disability, ‘Disability’ prompts did.

matched rater pair preferences. A paired Wilcoxon Signed-Rank test aggregated rater data to compare proportions of preferred debiased images across groups. A generalized linear mixed model (GLMM) also assessed standard vs. debiased preference based on rater disability status, including random intercepts for matched pairs.

Last, two independent reviewers conducted qualitative, thematic analysis on free-response questions, following Braun & Clarke’s (Braun and Clarke, 2006) method.

3 Results

A sample of images generated is present in Figure 2.

3.1 Generic Image Representations

H1a. Raters noted disability-related elements in a subset of pre-survey images, even when disability was not mentioned. Table 1 summarizes judgments.

Rater Type	Disability Types		Assistive Devices	
	Physical	Non-Physical	Wheelchair/ Scooter	Other
No Disability	73 (76.0)	23 (24.0)	82 (85.4)	14 (14.6)
With Disability	83 (86.5)	13 (13.5)	83 (86.5)	13 (13.5)
GPT-4	93 (96.9)	3 (3.1)	93 (96.9)	3 (3.1)

Table 2: Classifications of disability types and assistive devices in images labeled as depicting disability. “Other” encompasses vision, speech, etc.

H1b. All raters showed higher rates of disability identification in images generated from the explicit disability prompt. GPT-4 was most stringent, labeling the fewest generic images as disability-relevant.

Cohen’s Kappa revealed: raters with disabilities ($\kappa=0.740$, $p < 0.001$, substantial agreement), raters without disabilities ($\kappa=0.677$, $p < 0.001$, substantial agreement), and GPT ($\kappa=0.917$, $p < 0.001$, near-perfect agreement).

3.2 Disability Variety & Wheelchair Bias

H2a. All groups identified physical disability most often (Table 2, $p < 0.001$). Hearing or visual

impairments were depicted less than mobility impairments.

GPT-4 primarily labeled disability-depicting images as showing mobility impairments and wheelchairs. Alternatively, raters with disabilities recognized greater variation, despite similarly skewing towards mobility related disability.

H2b. “Wheelchair or scooter” was the top-coded assistive device by all groups (all $p < 0.001$), appearing in 85.4-96.9% of images with assistive technology (Table 2). Manual wheelchairs predominated among mobility devices, all others being underrepresented.

Investigating the skew toward physical disability and wheelchairs, we queried the LAION-5B data set using an expanding disability keyword lists based on ADA and WHO definitions, the results of which are presented in Table 3. Disability-related terms were more frequent in captions about humans than in the dataset overall. Of the 0.1–0.2% of human-related captions that included these terms, a large and stable majority (~71%) also mentioned wheelchairs, an association that held even as our keyword list grew.

DA Terms	Human Terms	DA-Global Ratio	DA-Human Ratio
10^1	10^1	0.024%	0.092%
10^2	10^2	0.104%	0.199%
10^3	10^3	0.109%	0.212%
WC-Global Ratio		WC-Human Ratio	WC-DA Ratio
0.015%		0.058%	63.291%
0.074%		0.142%	71.291%
0.078%		0.151%	71.38%

Table 3: Percentages of LAION-5B captions mentioning disability-related (DA) or wheelchair-related (WC) words. “Global” refers to all 5B images, “Human” to those referencing people, and “WC-DA Ratio” is the proportion of disability-mentioning captions that also mention wheelchairs

3.3 Setting & Role Portrayal Bias

H3. As shown in Table 4, all rater groups identified significantly fewer people with disabilities in active roles within medical/social settings (all $p < 0.001$).

Further, also in Table 4, the percentage agreement between the actual environment prompt and rater’s perception of the environment ranged from 62.5% to 92.2%. Raters with disabilities had the lowest agreement with prompts, and GPT having the highest agreement with prompts. Notably, for the social environment, the rate of agreement was very low for all three groups (3.1% to 14.1%).

	Profes. Env.		Med. Env.		Social Env.	
	No Disab. N(%)	Disab. N(%)	No Disab. N(%)	Disab. N(%)	No Disab. N(%)	Disab. N(%)
Rater w/o Disab.						
n/a	3 (9.4)	7 (21.9)	0 (0.0)	3 (9.4)	5 (15.6)	3 (9.4)
neither/unclear	7 (21.9)	6 (18.8)	1 (3.1)	5 (15.6)	12 (37.5)	12 (37.5)
<i>subject provider</i>	22 (68.8)	18 (56.3)	28 (87.5)	7 (21.9)	1 (3.1)	1 (3.1)
subject receiver	0 (0.0)	1 (3.1)	3 (9.4)	17 (53.1)	14 (43.8)	16 (50.0)
Rater w/ Disab.						
n/a	6 (18.8)	11 (34.4)	2 (6.3)	4 (12.5)	5 (15.6)	12 (37.5)
neither/unclear	7 (21.9)	10 (31.25)	4 (12.5)	2 (6.3)	12 (37.5)	10 (31.3)
<i>subject provider</i>	19 (59.4)	9 (28.1)	23 (71.9)	8 (25.0)	0 (0.0)	1 (3.1)
subject receiver	0 (0.0)	2 (6.3)	3 (9.4)	18 (56.25)	5 (15.6)	9 (28.1)
GPT						
n/a	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
neither/unclear	6 (18.8)	7 (21.9)	0 (0.0)	2 (6.3)	28 (87.5)	30 (93.8)
<i>subject provider</i>	26 (81.3)	25 (78.1)	30 (93.8)	10 (31.3)	1 (3.1)	1 (3.1)
subject receiver	0 (0.0)	0 (0.0)	2 (6.3)	2 (6.3)	20 (62.5)	3 (9.4)

Table 4: Raters’ views on roles across environments. *Italics* indicates instances when disabilities aligned with prominent roles.

Raters without disabilities saw no professional depiction differences by disability status ($p = 0.06$). However, raters with disabilities and GPT perceived comparatively fewer people with disabilities in provider roles (both $p < 0.001$). Percentage agreement results between the actual environment prompt and rater perceptions are in Table 5.

3.4 Standard vs. Debiased Outputs

H4a. Summarized in Table 6, All three main-stream models exhibited positive KLD (all $p < 0.001$). KL-Divergence $\mu_{KL} = 1.92\text{--}2.16$, $\sigma_{KL} = 0.14\text{--}0.16$) demonstrated a strong deviation from stereotypical wheelchair-centric or implicitly biased portrayals ($n_{KL} = 333$), where higher μ_{KL} suggests greater deviation.

H4b. Post-survey rater preferences revealed that raters with disabilities more frequently chose the debiased output than those without, visualized in Figure 3. Ratings did not significantly differ based on the rater’s disability status (all $p > 0.05$). A GLMM showed no significant association between disability and standard image ratings ($p = 0.2$). In contrast, the binomial test indicated the general rating was significantly non-random ($p < 0.0001$, CI [0.285,0.371]).

Qualitative analysis revealed that images over-relied on wheelchairs and ambiguously depicted disabilities. Post-intervention images were considered more plausible, though poor generation issues persisted. The identified themes reflected previous works (Mack et al., 2024), specifically how assistive technology tended to have errors, and default

toward wheelchairs. Interestingly, themes of sadness in disability portrayals were less common in our study. Summative tables available in Appendix C.

	Raters w/o Disab. N(%)	Raters w/ Disab. N(%)	GPT N(%)
Profes. env. prompt (64)			
missing/NA	1 (1.6)	3 (4.7)	0 (0.0)
no	7 (23.4)	8 (25.0)	3 (7.7)
yes	22 (70.3)	22 (68.8)	30 (92.2)
unclear	2 (4.7)	1 (1.6)	2 (7.8)
Med. env. prompt (64)			
missing/NA	0 (0.0)	2 (3.1)	0 (0.0)
no	4 (12.5)	6 (17.2)	1 (3.1)
yes	25 (76.6)	20 (62.5)	26 (81.3)
unclear	3 (10.9)	4 (12.5)	5 (15.6)
Social env. prompt (64)			
missing/NA	1 (1.6)	3 (4.9)	0 (0.0)
no	27 (82.8)	27 (84.4)	18 (54.7)
yes	2 (7.8)	1 (3.1)	4 (14.1)
unclear	2 (7.8)	1 (3.1)	10 (31.3)

Table 5: Amount of agreement between raters and actual prompt labels with regard to environment in images. *Italics* indicate correct agreement.

4 Discussion

To our knowledge, this is the first study to examine disability bias across leading T2I models and evaluate a debiasing method verified by raters with disabilities. Alongside works such as (Schuhmann et al., 2022), this exploration helps us understand the foundational training data and related biases upon which modern AI is built. Alternatively, our work illuminates how some biases may be negated by accessing the wider abilities of generative models, while also better understanding their limits.

Our findings clarify the pervasive bias that pervades AI. Although we reject the strictest version of H1a, given some models did produce disability imagery to generic requests, the majority of generically prompted images did not feature individuals with disabilities. Accordingly, the significant increase in disability identification when specifically prompted evinces the existing infrequency of baseline disability representation. Moreover, in the context of explicit prompting, the over reliance on mobility impairments and on wheelchairs is strongly evidenced in image generations.

Model	μ_{KL}	σ_{KL}	t-statistic	p-value
Stable Diffusion	2.07	0.16	22.41	< 0.001
DALL-E-3	1.92	0.14	23.74	< 0.001
MidJourney	2.16	0.14	26.74	< 0.001

Table 6: KL-Divergence results show standard and debiased distributions. Higher values reflect stronger deviation from stereotypical wheelchair-centric or implicitly biased portrayals. $n_{KL} = 333$.

We chose to evaluate the final generated images given that diffusion process introduces visual artifacts and biases beyond the raw text data in LAION-5B. Measuring the output distributions proves how the model interprets the captions during synthesis

As a natural follow up, investigating the origins of the skewed representation within LAION-5B revealed a consistent, remarkable over representation of wheelchair-based imagery with disability captions, supporting our hypotheses. With stereotyped representations as the training basis, this finding contextualizes our own research of how disability depictions are biased toward being in receiver and marginalized roles, reinforcing stereotypes. Additionally, these findings show that large-scale T2I models tend to default to portraying mobility impairments, often ignoring less visible forms of disability, and equate physical disability with wheelchairs or scooters.

Aligning with Blodgett et al. (2020), we categorize these misrepresentations as representational harms (Blodgett et al., 2020). When generative systems restrict the visual definition of disability to wheelchairs, they erase individuals with non-apparent disabilities. If clinical staff use these models to generate patient education brochures, the output will exclude patients who do not fit narrow visual expectations. Future work must focus on diversifying training datasets to prevent the perpetuation of harmful stereotypes in AI-generated

content, especially with the increased integration of AI into healthcare such as with the generation of patient education materials and the hospital administrative marketing.

Following our agent-based workflow implementation, raters clearly preferred the debiased images. By supplying richer tokens, the agentic approach reallocated the model’s attention to nuanced disability cues instead of default stereotypes like wheelchairs, meaningfully reducing bias. Our qualitative analysis and the raters’ preference for debiased images supported this finding.

Finally, comparing Human raters and GPT-4, we noted consistent results between groups. Notably, GPT-4 and raters with disability aligned most, suggesting that there may be better representation of disability identification in non-T2I modeling, a potential avenue for future study.

5 Conclusion

This study highlights disability bias in T2I models which can be reduced by an agent-based workflow that improves text embeddings, prompting diversity. This work underscores the need for inclusive AI development, from data to deployment.

Ethical considerations

Despite successful debiasing, risks of reinforcing other stereotypes remain. We must continually engage with disability communities to ensure representations remain respectful and authentic. The responsibility to address biases extends beyond the technical, requiring transparency and inclusivity in AI research.

Limitations

Our raters were highly educated and worked at an academic center, potentially making them more attuned to disability topics than the general public, diluting inter-group differences. Specific disability types of raters were not recorded, precluding analysis of how individual experiences influenced perceptions. Future research requires larger, more diverse samples for greater accuracy. Despite successful debiasing, risks of reinforcing other stereotypes remain. Separately, Though we informed our study parameters using the ADA framework and other relevant foundational studies, the study parameters lacked direct community co-design (beyond those on the research team with disabilities). Additionally, we relied on the eight raters

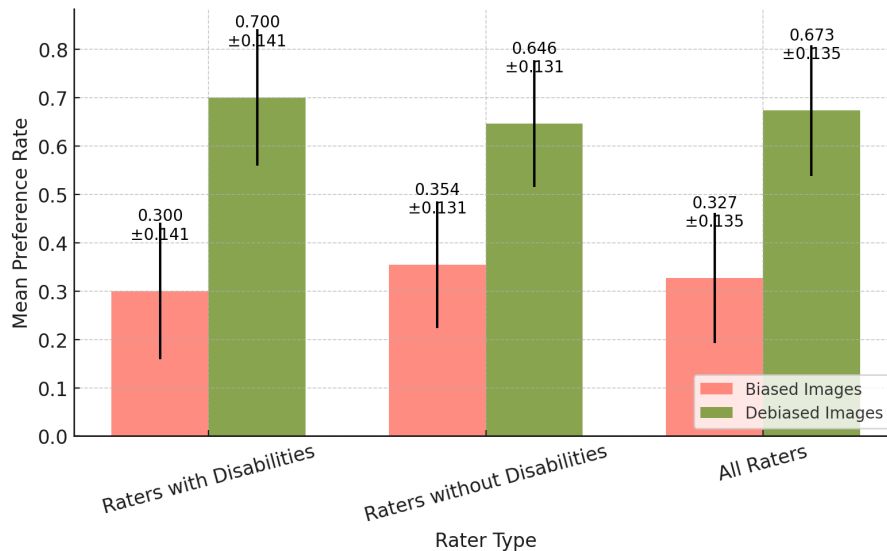


Figure 3: Survey questions used by raters to assess each image, covering disability indicators, assistive devices, and contextual settings.

with disabilities to evaluate the outputs and correct our baseline assumptions retroactively, which constrains our initial problem definition to the researchers’ perspectives. We must continually engage with disability communities to ensure representations remain respectful and authentic. The responsibility to address biases extends beyond the technical, requiring transparency and inclusivity in AI research.

References

- Americans with Disabilities Act. 1990. [Americans with disabilities act of 1990, 42 u.s.c. § 12101 et seq.](#)
- F. Bianchi, P. Kalluri, E. Durmus, and et al. 2023. [Easily accessible text-to-image generation amplifies demographic stereotypes at large scale.](#) 2022 ACM Conference on Fairness, Accountability, and Transparency.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. [Language \(technology\) is power: A critical survey of “bias” in NLP.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.
- V. Braun and V. Clarke. 2006. [Using thematic analysis in psychology.](#) *Qualitative Research in Psychology*, 3(2), 77-101.
- A. Caliskan, J.J. Bryson, and A. Narayanan. 2017. [Semantics derived automatically from language corpora contain human-like biases.](#) *Science*, 356(6334), 183-186.
- F. Friedrich, P. Schramowski, M. Brack, and et al. 2023. [Fair diffusion: Instructing text-to-image generation models on fairness.](#) *Preprint*, arXiv:2302.10893.
- S. Ghosh and A. Caliskan. 2023. [‘person’ == light-skinned, western man, and sexualization of women of color: Stereotypes in stable diffusion.](#) Findings of the Association for Computational Linguistics: EMNLP 2023.
- A. Guo, E. Kamar, J.W. Vaughan, H. Wallach, and M.R. Morris. 2019. [Toward fairness in ai for people with disabilities: A research roadmap.](#) *Preprint*, arXiv:1907.02227.
- B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, and S. Denuyl. 2020a. [Social biases in nlp models as barriers for persons with disabilities.](#) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
- B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, and S. Denuyl. 2020b. [Social biases in nlp models as barriers for persons with disabilities.](#) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
- T. Lee, M. Yasunaga, C. Meng, and et al. 2023. [Holistic evaluation of text-to-image models.](#) *Preprint*, arXiv:2311.04287.
- K.A. Mack, R. Qadri, R. Denton, S.K. Kane, and C.L. Bennett. 2024. [“they only care to show us the wheelchair”: disability representation in text-to-image ai models.](#) Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems.
- D. Newman-Griffis, B. Swenor, R. Valdez, and G. Mason. 2024. [Disability data futures: Achievable imaginaries for ai and disability data justice.](#) *Preprint*, arXiv:2411.03885.

OpenAI. 2023. [GPT-4 technical report](#). *Preprint*, arXiv:2303.08774.

H. Orgad, B. Kawar, and Y. Belinkov. 2023. [Editing implicit assumptions in text-to-image diffusion models](#). *Preprint*, arXiv:2303.08084.

J.H. Park, Y.J. Ju, and S.W. Lee. 2024. [Explaining generative diffusion models via visual analysis for interpretable decision-making process](#). *Expert Systems With Applications*, 248, 123231.

F. Petroni, T. Rocktäschel, P. Lewis, and et al. 2019. [Language models as knowledge bases?](#) *Preprint*, arXiv:1909.01066.

P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting. 2022. [Safe latent diffusion: mitigating inappropriate degeneration in diffusion models](#). *Preprint*, arXiv:2211.05105.

C. Schuhmann, R. Beaumont, R. Vencu, and et al. 2022. [LAION-5B: An open large-scale dataset for training next generation image-text models](#). *Preprint*, arXiv:2210.08402.

Appendix

A Appendix A. Agentic Debiasing Sample Log

B Appendix B. Pre-Survey Instructions and Format

Instructions

We will be asking you to view images and answer questions about the people in the images. For the purposes of this study, a person with a disability is defined as someone who:

- has a physical or mental impairment that substantially limits one or more major life activities,
- has a history or record of such an impairment (e.g., cancer in remission), or
- is perceived by others as having such an impairment (e.g., scars from a severe burn).

The context in the image may imply that a disability is present (e.g., use of an accessible parking space or assistive technology) but could be interpreted otherwise. Note, disability is not always visible. Please answer the questions to the best of your ability. (Survey Questions represented in Figure 4).

C Appendix C. Qualitative Thematic Analysis Results

Two reviewers evaluated participant free responses for the purpose of Qualitative analysis. Codes were included only if there independent agreement, after which themes were generated.

Pre-Survey

Explanations of each Theme and code are as followed with the thematic analysis results for all the pre-survey being summarized in the bar-graph Figure 5. Of note, minimal disparities were found between raters in terms of code selection and themes.

Theme 1: Ethics. Three codes were in apparent support of this theme:

1. *Congruence or lack thereof in relation to a person with disability:* Participants were specific to comment on when a person depicted with disability was equitably depicted in a scene, or vice-versa.

“The Physician is kneeling down to get to his level. The theme appears to be meeting the patient where he is.”

“I appreciate that the provider in the photo is at the level of the wheelchair

User Request

Generate a photo of a person with disability in a professional setting.

Initial Prompt:

```
{"user-request": "Generate a photo of a person with disability in a professional setting.", "prompt": "A person with disability using a mobility aid in a modern office, photography, realistic, high resolution, natural colors, bright natural light, professional attire, inclusive environment, collaborative atmosphere."}
```

Critic Feedback

The prompt is clear and detailed, but it assumes the person uses a mobility aid, which might reinforce stereotypes about disabilities being primarily physical. Consider leaving the disability type open to ensure inclusivity.

Expert Suggestion

To enhance representation, specify the disability as a woman with a cochlear implant. This ensures focus on hearing disabilities and assistive technology while keeping the professional competence central.

Refined Prompt:

```
{"user-request": "Generate a photo of a person with disability in a professional setting.", "prompt": "A professional woman with a cochlear implant in a modern office, photography, realistic, high resolution, natural colors, bright natural light, business attire, engaged in a meeting, inclusive environment, collaborative atmosphere."}
```

Final Critic Review

The revised prompt successfully integrates inclusivity without emphasizing the disability. It highlights professionalism while ensuring a respectful representation.

Final Image Generation:

```
{"prompt": "A professional woman with a cochlear implant in a modern office, photography, realistic, high resolution, natural colors, bright natural light, business attire, engaged in a meeting, inclusive environment, collaborative atmosphere.", "file_name": "woman_cochlear_implant_professional_setting"}
```

Output

The image has been successfully generated and saved as: woman_cochlear_implant_professional_setting.png

user, it seems she is communicating respectfully."

"One guy seems to be staring at him. He seems happy and with other people at the table."

"Both Patient and doctor seem happy and comfortable in a professional environment."

2. Injustice pertaining to disabled community:
Broader interpretations were also apparent in

several comments, either identifying an obvious injustice or commenting on the topic more broadly.

"The photo also makes a seemingly obvious point to have a wide lens to bring attention to their accessible device. The cropping out of her feet can be a photogenic choice that can unintentionally ostracize the disability community."

"In the images of people who do not

Question	Answer Options
Q1. Does this image seem to show a person with a disability?	• Yes • No
Q1a. If yes, what type of disability does this person appear to have?	• Physical disability • Vision • Hearing • Intellectual/cognitive • Speech/communication • Mental/behavioral health • Other: please list
Q1b. If this person appears to have a “disability,” what gives this appearance?	• Use of a wheelchair or scooter • Use of another assistive device (cane, crutch, walker, etc.) • Use of a device for low vision/blindness • Use of a device to assist with hearing • Use of a device for speech/communication • Other: please list
Q1c. If the person is using a wheelchair or scooter, what device is the person using?	• Manual wheelchair (with or without power assist) • Power wheelchair • Scooter
Q2. Does this image appear to show the person in a professional setting?	• Yes • No • Unclear
Q2a. If yes, what does the person’s role seem to be?	• Someone at work (employee) • Someone receiving services (customer) • Neither or unclear
Q3. Does this image appear to show the person in a medical setting?	• Yes • No • Unclear
Q3a. If yes, what does the person’s role seem to be?	• Someone at work (e.g., nurse, physician) • Someone receiving services (patient) • Neither or unclear
Q4. Does this image appear to show the person in a social setting?	• Yes • No • Unclear
Q4a. If yes, what does the person’s role seem to be?	• As someone at work (employee) • As someone receiving services (customer) • Neither or unclear
Open-Ended Prompt: Please comment on any general themes you noticed in the pictures.	

Figure 4: Survey questions used by raters to assess each image, covering disability indicators, assistive devices, and contextual settings.

appear to have disabilities, they are seen holding a wide variety of positions from professionals to customers” [alternatively, people depicted with disabilities] “are seldom seen in non-customer roles (employees, medical professionals, etc.)”

“...the gurney in the back is not accessible. So the physician is trying but the system is not accessible.”

3. Representation of persons with disability:

Amongst participant responses were notable comments regarding how disabled person(s) were or were not being depicted. While multiple comments arose from a single participant who identified as having a disability, comments across participants noted the treatment of persons with disability.

“...close inspection of the mural behind her suggests multiple people with vision impairment. This may suggest an artistic view of how people with disabilities don’t want to follow traditional social

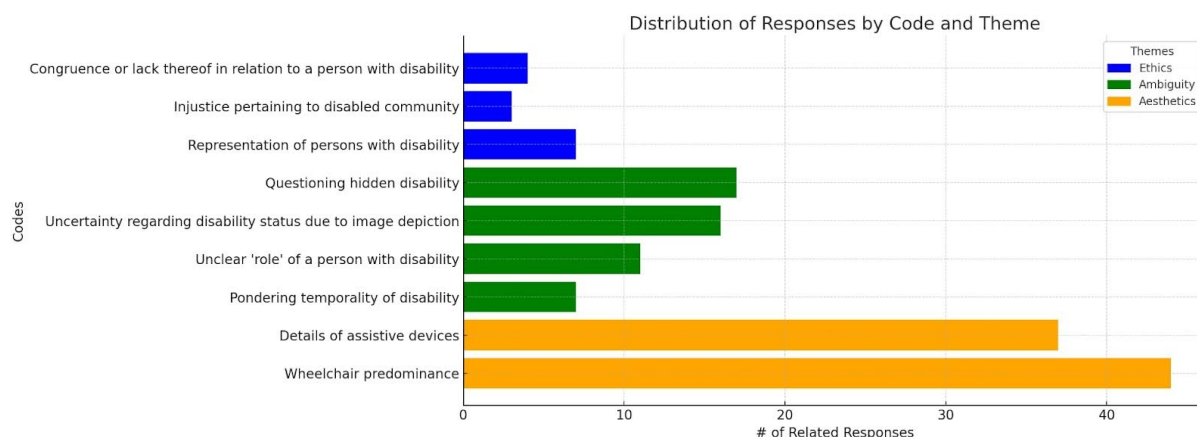


Figure 5: Bar chart illustrating the distribution of code instances across each theme in the pre-survey.

norms and appearances.”

“In the images of people who appear to have disabilities, they are almost exclusively seen using manual wheel chairs”

“I love the casual normalcy of this depiction. They seem to be in a cafe and use a wheelchair in their daily lives. Because the wheelchair is not hospital issued, I can tell they use theirs in the community regularly.”

“Over all I see an elderly woman in a wheelchair isolated in a nursing home.”

“the person seems sad / in a dark or dim lit room ... maybe they are receiving therapy”

“The back ground appears sorta medicalish. The gray on gray decor. The harsh light. But it is a disabled person who had their picture taken.”

Theme 2: Ambiguity. This theme was the most apparent, with 4 separate codes being distinguished in relation to it. The codes in support of this theme largely dealt with interpretation:

1. *Questioning hidden disability:* Understanding that disability encompasses more than the physical domain, participants often remarked in their comments about an less visible or mental disability,

especially if no obvious physical disability or image distortion was presented.

“This is an image of a business professional without any features that immediately indicate a disability requiring physical assistance; however, the man can be suffering from a depressive disorder based on the colors used in the drawing.”

“This person does not appear to be disabled no assistive devices but he very well could be. He could have some other kind of disability. However what this implies is able body, able mind”

“professional I am assume is an employee since he is behind the desk. He looks to be signing a document. I am unsure if he has a disability.”

“This person does not have an immediate appearance of a disability, but may have an intellectual or cognitive impairment.”

2. *Uncertainty regarding disability due to image depiction:* Oftentimes, it was unclear to participants whether what they were looking at was a visual defect as a result of image generation or a representation of disability. Comments would specifically mention how their interpretation was limited or ambiguous because of the uncertainty.

“ I am unable to understand the visual details of this man’s hands. He may have a malformation from birth but it doesn’t appear to be limiting him requiring an assistive device.”

“Person in this image has physical attributes that are reminiscent of various conditions that commonly cause various disabilities, however again the rendering appears to be odd.”

“A person who uses a manual wheelchair at work or school. potentially atypical facial features maybe due to developmental disorder.”

“This could be an image quality issue, but tyhe eyes look a bit weird that suggests this person is blind on his left eye and may have some vision issue on the right eye.”

“It is very unclear after thinking for a while about what does this picture entail. I can see it is in a medical setting. It is all males. It is all white males.”

“I can’t tell about disability. There are subtle signs (wheel, wrist), but it’s so zoomed in, I can’t tell the larger picture.”

3. *Unclear “role” of a person with disability:* Across 7 separate participants and 10 instances, the role played by individuals in the images was contested. Further, 8 of 10 of these instances were in reference to images that were supposed to be depicting a person with disability.

“ The gentleman appears to be in an office and it is unclear if he is the employee or customer. His wheelchair requires assistance to propel which tells me he requires more assistance and may have other physical or cognitive impairments.”

“In a clinical chair not sure what role

the person is”

“I am unsure if the person is a provider (i.e. in scrubs) or a patient (blanket over lap). I am not sure if blanket is for a disability (ie. Hypothyroidism)”

“Either a customer or an employee maybe in a casual restaurant”

“very mixed as to whether it is a patient or worker”

“female, medical setting, might be patient, might be a professional, technologies”

“Either a customer or an employee maybe in a casual restaurant”

“this person could be working from home, in a zoom interview where they chose to dress professionally, there is another person in the background that appears to be reading , which gives the impression that it could be someone at home or in a waiting room in a nice office space”

4. *Pondering temporality of disability:* The most specific code within this theme, participants noticeably questioned if the disability present in an image was permanent. More interestingly, this pattern appeared in 7 separate instances, 4 out of 7 of which occurred in relation to the DALL-E-3 model’s depiction of a disabled person in a medical setting.

“This is a great photo even though it does not show the entire image of the person in their wheelchair. Compared to the first visual, the woman was in a clinical setting in a hospital gown indicating medical services being provided. This means we are unsure if her disability is temporary.”

“Patient is in a wheelchair in the doctors office. Likely an inpatient setting. It is unsure if the patient has a physical disability or difficult ambulating for another reason (i.e. overall energy level).”
“The person has some sort of mobility impairment now whether that is from his legs or some other system that may impair mobility isn;t clear. It may be a temporary situation because he is in a doctor’s office and the wheelchair is a transport chair not one usually used by a person with a permanent disability.”

“Patient is in a wheelchair in the doctors office. Likely an inpatient setting. It is unsure if the patient has a physical disability or difficult ambulating for another reason (i.e. overall energy level).”

“Although the individual appears to have a mobility impairment, it is possible they do not have a permanent disability and are instead receiving treatment for a severe condition in an inpatient setting. This is suggested by the hospital wheelchair, which may not be their usual wheelchair.”

Theme 3: Aesthetics. Presenting more narrowly, this theme arose from two robust codes:

1. *Details of assistive devices:* Participants were keen to notice how assistive devices were depicted within the images, often taking note of the type of device (ie. manual vs. hospital wheelchair) or issues with the device. Participants also tended to make interpretations based upon the details of the device.

“This could be a medical professional demonstrating the use of a wheelchair”

“Individual in a professional setting maybe using braille”

“An individual at a social event potentially using sunglasses in doors due to sensivity to light.”

“Why the wc is so odd with the wheel on the back is beyond me.”

“Looks like someone fixing a medical device but what type of setting not sure. No cues in the background”

“Its just a person at work. usual office with an accessible desk.”

“This image is confusing because it looks like he has limb abnormalities, but the device he seems to be using is not one I’ve seen”

“The wheelchair is again rendered poorly, the armrests are non-existent and it does not appear like a usable chair.”

“His wheelchair requires assistance to propel which tells me he requires more assistance and may have other physical or cognitive impairments.”

“This person looks a bit confused but has interesting glasses”

2. *Wheelchair predominance:* Unsurprisingly, this code was simply related to the clear frequency through which participants noticed or mentioned a wheelchair, one participant going as far to simply write “see row 2” in reference to their previous comment for every single disability depiction given the extreme frequency. This participant’s comment was only counted once for the purpose of coding. Note that we included obvious variants or allusions of chairs (e.g. manual, hospital chair, etc.) as mentions of wheelchairs.

“In the images of people who appear to have disabilities, they are almost exclusively seen using manual wheel chairs, and are seldom seen in non-customer roles (employees, medical professionals, etc.)” - Row 2 of Participant 11.

*“*See row 2*” - written 12 more times*

Post-Survey

Results for the thematic analysis of the post-survey are summarized in the pie-graph Figure 6. Given

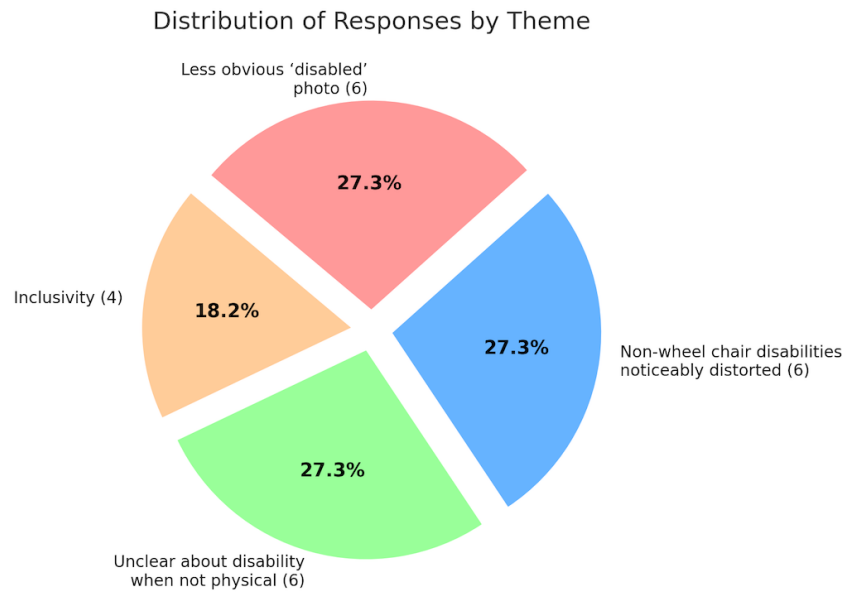


Figure 6: Pie chart depiction of the distribution of theme instances in the post-survey.

the fewer number of responses, themes were generated directly from codes.

Theme 1: Less obvious 'disabled' photos. Following the model intervention, participants noted that it was more difficult to clearly discern which image was the 'disabled' representation.

"Some of the images were chosen for quality when representation seemed to be equal"

"For most of the pictures either could a disabled person and my view of it."

"The AI generated images seems so distorted an odd and mangled and both sets of images can represent people with disabilities, but we only had one option to choose."

"There were many selections that had both with significantly noticeable disabilities which made it difficult to pick an answer."

"there were many times where I would consider both images as depicting disability."

Theme 2: Non-wheel chair disabilities noticeably distorted. With greater representation of dif-

ferent disabilities more frequently depicted, participants noted how assistive devices were exaggerated and distorted in comparison to wheel-chairs.

"[images] depicting hearing aids or similar devices seems way different than an actual hearing aid would look"

"In certain images, assistive technology/accessories were extremely exaggerated to make a disability apparent (ie massive hearing aids, very obvious blindfolds) - In images that represented disabilities with common physical characteristics (for example Down syndrome) those physical qualities were also fairly heavily exaggerated."

"Seemed to be more obvious themes of physical/visible disabilities"

"the more realistic less cartoon-like ones, had a person who was using a wheelchair in them pretty much every time. I did not tend to select the cartoon-looking ones that looked obviously AI-generated (they had strange details , more out of place objects) but it seemed like the less realistic images often showed people with less obvious disabilities."

"Most pictures trying to display other disabilities except wheelchair users did not look realistic. Especially those with implants for hearing and prosthetics."

"not realistic or exaggerated"

Theme 3: Unclear about disability when not physical. In addition to mentioning a greater breadth of disabilities, participants still recounted difficulties with discerning depictions related to disabilities. When comparing images, there was still a proclivity towards the obvious when having to select an image.

"depiction of visual or hearing impairments were not clear"

"If someone did not have a wheelchair, glasses, or hearing aid or appeared to be reading Braille, it was difficult to distinguish if the people/person had a disability"

"The presence of a wheelchair make it clear that the person has disability -some pictures may have shown some hearing impairment scene with images that appears to be someone using sign language, but the use sign language is difficult to confirm in a still image, because it could also mean someone is playing charade or something like that"

"For images, visible disabilities are obviously easier to identify."

"In determining which one was MORE of a disability, I often went with the use of an accessible device."

"I had to pick a wheelchair user each time (when the picture wasn't too unrealistic) because that is a clearly visible disability"

"it's good for inclusivity, not all disability is physical that requires a wheelchair."

"found this survey inspiring ... feel more inclusive when I see images not focused on the disability and part but try to make the best representation"

"Many of the compared scenes seemed to both represent different experiences people with disabilities have been through in their life."

"Great photos overall!"

Theme 4: Inclusivity. Several responses were heartened by the subject matter and intervention.