Position: To Make Text-to-Image Models that Work for Marginalized Communities, We Need New Measurement Practices for the Long Tail

Nari Johnson[1,2*] **Hamna**[1] **Deepthi Sudharsan**[1] **Theo Holroyd**[3]
Samantha Dalal[1,4] **Siobhan Mackenzie Hall**[1,5] **Jennifer Wortman Vaughan**[1]
Daniela Massiceti[1] **Cecily Morrison**[1]
[1]Microsoft Research [2]Carnegie Mellon University [3]Independent Researcher
[4]University of Colorado Boulder [5]University of Oxford

Abstract

While the capabilities of frontier text-to-image models are rapidly improving, they often fail to represent the low data, long tail concepts that matter to historically marginalized communities. Effective measurement is a critical first step towards identifying and addressing these errors, yet little work has validated if existing T2I evaluation metrics work for the long tail. In this paper, we draw upon two community-based case studies to identify challenges with applying best practices to validate T2I metrics using human preference data. We show that available approaches to create and validate evaluation metrics break down when applied to tail concepts because of the need for community knowledge (scaling community annotations) and challenges achieving a range of good and bad images (shades of bad). We take the position that methodological innovation is needed to develop measurement practices that work for the long tail. We outline directions for future work that moves beyond traditional approaches to measurement towards imagining new ways to center community expertise throughout the measurement process.

1 Introduction

The capabilities of today's generative AI models are stunning, but they do not work for many marginalized communities [9, 24, 25, 30, 57, 82]. A growing body of work has surfaced how text-to-image ("T2I") models can contribute to errors of representation that can be upsetting to view, perpetuate stereotypes, or result in cultural erasure [13, 42, 45, 59, 73]. To name a few examples, state-of-the-art T2I models often fail to generate accurate and dignified depictions of people with disabilities [57], African dishes [59], or scenes of everyday life in South Asia [72]. Research has shown that despite their cultural significance to communities, many cultural artifacts are systematically excluded from large web-scraped pretraining datasets [31, 37, 60, 63] and consequently end up in the "long tail" of machine learning [85, 100].

Measurement is critical to identifying and addressing errors of representation for marginalized communities. Effective measures enable practitioners to compare and select models [32, 90] and to "guide the development of AI systems themselves" [90], e.g., by using the measure as a reward model [48, 93]. Not least, valid measures help to reveal underperformance and the progress being made toward addressing it. For these reasons, there are increasing calls to develop a "mature evaluation science for generative AI systems" [91], recognizing that commonly used "one size fits all" benchmarks fail to account for a range of factors that matter to those who use these models [75, 87, 91].

[*] Work completed as an intern at Microsoft Research. Correspondence to: Nari Johnson, narij@andrew.cmu.edu

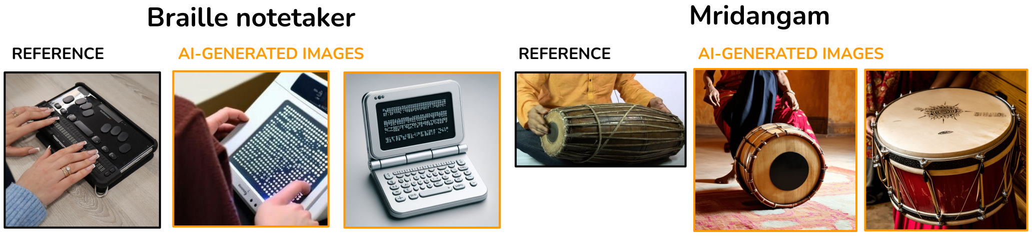

Figure 1: **State-of-the-art text-to-image models fail to generate accurate depictions of artifacts from marginalized communities.** Reference photos and AI-generated images (from Stable Diffusion 3 [22] and DALLE-3) of two cultural artifacts. (Left) A braille notetaker is a small portable electronic device that provides braille users with a means to write and read digital text. (Right) A Mridangam is a popular South Indian percussion instrument from Tamil Nadu. Alt text is available in Appendix A.

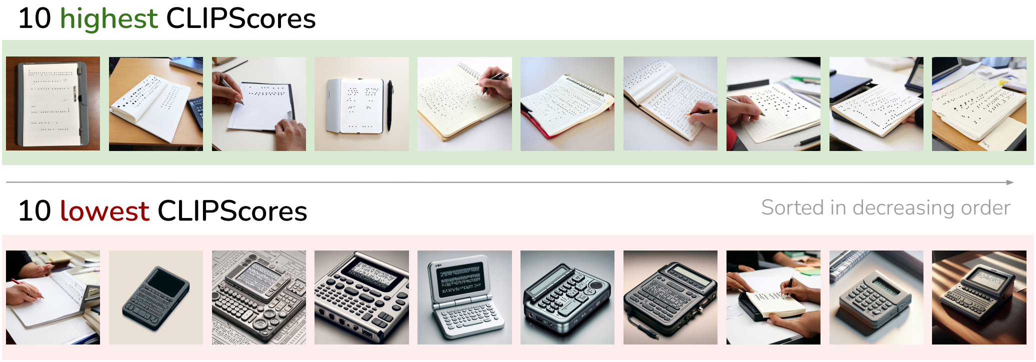

Figure 2: **CLIPScore favors inaccurate and offensive depictions of a braille notetaker.** Generated images of a braille notetaker that had the 10 highest and lowest CLIPScores. Images were generated by Stable Diffusion 3 and DALLE-3 (Appendix B). Images that depict notetaking using ink on paper are assigned higher scores than images that depict notetaking using an electronic device. An independent two-sample t-test using the full sample of 60 images produced a statistically significant difference in the CLIPScores assigned to the two groups ($p < 0.01$). Community members found depictions that used paper to be the most inaccurate and offensive, indicating misalignment with the metric.

One important dimension of T2I evaluation is measuring image-prompt alignment (sometimes called consistency or faithfulness): the extent to which a generated image matches its input prompt [51, 77, 80]. Popular image-prompt alignment metrics increasingly rely on secondary large pretrained vision-language models to score T2I outputs [77, 80] — part of a larger trend towards reliance on secondary models throughout model development and evaluation [44]. Despite their wide usage, little work has been done to understand how well these metrics work when applied to depictions of the people, places, or things that are bound up with marginalized communities' culture or identity [45, 81]. This lack of scrutiny is especially concerning given that large pretrained models have been shown to encode biases against, or lack knowledge about, underrepresented groups [60].

As an illustrative example, we compared a set of AI generated images based on CLIPScore, a popular T2I evaluation metric that uses OpenAI's CLIP model [34, 51, 74]. Figure 2 displays the ten highest- and lowest-scoring images for the prompt "braille notetaker," an electronic device commonly used by blind people to take notes [78]. The CLIPScores display a clear trend: images showing notetaking with ink on paper are consistently assigned higher scores than those showing an electronic device. However, ink-on-paper depictions were consistently ranked as *least preferred* by blind community members, who found these depictions inappropriate because ink is not tactile, and cannot be read by someone who is blind. This example shows how without proper validation, available T2I metrics may mistakenly favor representations of tail concepts that are inaccurate or even offensive.

Yet, current approaches to validating T2I metrics for the long tail are problematic. To argue this point, we draw on a project researching community-engaged T2I evaluation of cultural artifacts with two marginalized communities: (1) members of the blind and low vision community in the United Kingdom, and (2) residents of two South Indian nation-states: Tamil Nadu and Kerala. Following existing approaches to validate metrics by correlating metric scores with human preferences, we uncover two significant methodological challenges that arise in marginalized contexts (Section 4). First, we struggled to "scale community annotations" (Section 4.1) to obtain the sample size needed to make statistically meaningful claims about metric performance. Second, we found that frontier models failed to produce accurate depictions of tail concepts, forcing annotators to rate and choose between different types of errors, or "shades of bad" (Section 4.2).

Together, these challenges illustrate that validating a metric for the long-tailed context of marginalized communities, let alone using such a metric to drive model improvement, remains an open challenge. Therefore, **we take the position that we need new measurement practices to push the frontiers of representation in T2I models for marginalized communities**. We use our case studies to discuss how future work can investigate and address the challenges that we surface within existing frameworks for metric design and validation. We emphasize the opportunity for future work that moves beyond the status quo of how measurement is currently done, to imagine new ways to center community expertise throughout the measurement process.

2 Background & Related Work

2.1 Existing approaches to T2I evaluation

Our position contributes to a growing number of studies examining how T2I models represent marginalized communities and cultures [9, 18, 31, 73, 81, 89]. By inviting community members to respond to AI-generated images, qualitative studies have shaped how researchers understand representational harms in specific contexts [57, 61, 72]. Other efforts aim to make existing harms legible by crowdsourcing datasets of input prompts so that future evaluations can include content that matters to marginalized communities [42, 59, 69, 82, 102]. However, even when representational errors are qualitatively surfaced (e.g., frontier models often produce "errors in rendering assistive technologies" [57]), only a small subset of this work includes *quantitative* evaluations that propose *operational measures* that can be used to score image outputs (e.g., a metric that can detect if an assistive technology is rendered inaccurately in an image) [18, 31, 87].

To measure model performance, practitioners often default to using available image-prompt alignment metrics [32, 33, 87]. Given a model input x (e.g., a text prompt) in set $\mathcal{X}$ of all possible inputs, and a model output y (e.g., a generated image) in set $\mathcal{Y}$, an evaluation metric $\mu : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ maps the input-output pair to a score that reflects a construct of interest (e.g., image-prompt alignment). Saxon et al. [80] groups existing image-prompt alignment metrics into three classes: (1) methods that compare image and prompt embeddings (such as CLIPScore [51]), (2) VQA-based approaches that check if a generated image y meets a set of requirements derived from input prompt x (such as TIFA [38]), and (3) caption-based approaches that first generate captions for images y, which are compared to input prompts x (such as LLMScore [55]). Often conceived as a more efficient and scalable alternative to reference-based metrics [34], these reference-free metrics make use of "the relationships learned by pretrained vision-language models" [34] to score generated images.

However, recent work suggests that these metrics can replicate the biases of underlying models, limiting their effectiveness for marginalized groups [29, 49, 50, 94]. For example, in the context of image captioning, Kreiss et al. [49] found that CLIPSscores did not align with blind and low-vision users' preferences between captions. Thus, uncritical application of existing metrics can obfuscate underperformance or even lead practitioners to select models that perpetuate exclusion for marginalized groups [21, 32]. Recognizing the potential limitations of widely-used metrics, in this work we contribute an understanding of how existing measures might be *validated* to see if they are appropriate for evaluating depictions of concepts that matter to a community.

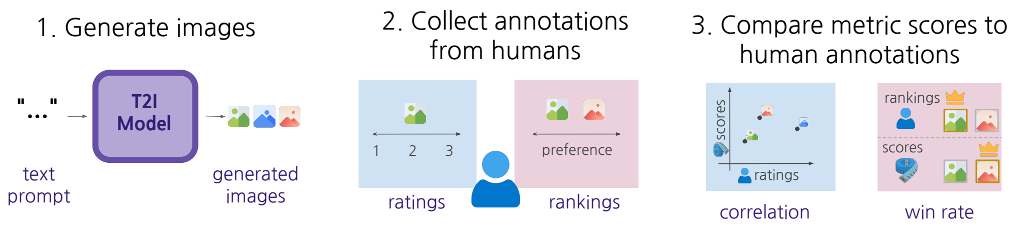

Figure 3: **Existing approaches to T2I metric validation.** An overview of typical metric validation studies that use human preference data in three steps. (Step 1) The study designer uses text prompts and a T2I model to generate a dataset of images. (Step 2) The study designer shows the generated images to an annotator, who provides annotations in the form of ratings (numeric scores) or rankings (ordered preferences between generated images). (Step 3) The study designer compares metric scores to human annotations by running a correlation test (for ratings), or by calculating whether the metric preferences the same images as humans (the metric's "win rate", for rankings).

2.2 Validating T2I evaluation metrics

The dominant approach to validating measures for T2I models involves crowdsourcing "true preference" data from human annotators to compare to metric scores [21, 38, 48, 64]. As shown in Figure 3, a typical workflow involves: 1) generating images using a set of prompts and chosen models; 2) developing task(s) for collecting human annotations; and, 3) running statistical tests to compare the collected preference data (taken as gold standard) to metric scores in order to validate the metric. Intuitively, a metric is valid if it is empirically aligned with human preferences, assigning higher scores to text-image pairs that annotators prefer.

Human preference annotation tasks (Step 2) can take two possible forms: rating tasks and ranking tasks [64]. Rating tasks instruct annotators to assess the quality of individual text-image pairs (x, y), such as asking "On a scale of 1 to 5, how well does the image y match the description x?" [51, 64]. Given a dataset of n text-image pairs $\{(x_i, y_i)\}_{i=1}^n$, metric scores $\{s_i := \mu(x_i, y_i)\}_{i=1}^n$, and average ratings $\{r_i\}_{i=1}^n$, the metric is typically validated by reporting the rank correlation (e.g., using Spearman's Rho or Kendall's Tau) between the human ratings $\{r_i\}$ and metric scores $\{s_i\}$ [34, 38, 64] (Figure 3, Right). The resulting correlation captures the extent to which the generated images with higher ratings are assigned higher scores by the metric.

In contrast, ranking tasks ask annotators to make comparative judgments between multiple outputs, such as asking "Which of these two images (y or y') best matches the prompt x?" [79, 97]. Other ranking tasks ask annotators to rank 3 or more images in order of most to least preferred [93]. For a fixed input prompt x and outputs y, y', the metric "wins" if it assigns a higher score to the image that the majority of annotators preferred, $p(y, y') \in \{y, y'\}$:

$$\mathbb{1}\left(\mu(x, y) \geq \mu(x, y') \iff p(y, y') = y\right) \tag{1}$$

The "win rate" metric averages these indicator variables $\mathbb{1}(\cdot)$ across the dataset of input prompts (Figure 3, Right). Thus, the win rate captures how frequently the metric favors the generated images that are preferred by the majority of annotators.

Prior studies that validate T2I metrics use common datasets of prompts such as captions written for real photographs from MS-COCO [52], handwritten prompts designed to assess specific T2I capabilities (e.g., spatial reasoning) [79, 80, 97], or "in-the-wild" prompts written by T2I users [48]. While handwritten prompt datasets aim to broaden evaluation beyond the "common scenarios" captured in MS-COCO captions [79], existing efforts to make prompt datasets more "challenging" prioritize compositional reasoning (e.g., "a kangaroo in a blue hoodie") over the inclusion of long-tail concepts or marginalized cultures [39, 79, 80, 97]. In this paper, we join calls for "contextual metric meta-evaluation" [21] in examining the performance of evaluation metrics *within* specific contexts. Specifically, we explore how to validate whether metrics are meaningful and useful measures for depictions of the tail concepts that matter to marginalized communities.

2.3 Generative AI evaluation as a human crowdsourcing task

Work on human evaluations of generative AI outputs, while nascent, shares many of the long-studied challenges of other crowdsourcing or collaborative computing tasks [2, 3, 20, 88]. A growing body of literature has established how annotators of AI outputs often disagree with each other, posing challenges for obtaining "gold" labels of a construct of interest [4, 27, 28, 36]. Variance across annotators can come from many underlying causes that emerge from relationships between the *annotator*, the *construct* being measured (by the annotation task), and the *media* being annotated [67]. Scholars have shown that disagreement is common in tasks where annotators are asked to make subjective judgments [2, 27, 47], such as whether a generated image is "harmful" [46, 76]. Variance can also arise when different annotators hold varying levels of prerequisite knowledge about the media being annotated, even for seemingly "objective" tasks such as recognizing whether a photograph contains a species of bird [67]. To address this variance, scholars have developed several best practices, such as using multiple annotators to report disagreement and reduce noise [36, 64, 70].

While many AI evaluations engage human annotators as "anonymized crowdworkers" [64], several researchers have advocated for alternative models of recruitment and participation in generative AI evaluation [8, 31, 43, 73, 84]. Recent studies reveal the critical role that annotators' identities and interpretations play in shaping how data is annotated [35, 43, 73]. For example, Hall et al. [30] found that annotators who live outside of a country are more likely to rate exaggerated, stereotypical depictions of the region as being "representative." In contrast, annotators who live in that region can draw upon on their experiences to provide more accurate annotations. This illustrates the value in directly engaging with community members to evaluate AI images. Our work draws upon these perspectives from past work to highlight the role of community knowledge in validating measures.

3 Case Studies

Our research team conducted a series of workshops with members of two marginalized communities that past work has shown are represented poorly by T2I models: members of the blind and low vision community in the UK [57], and residents of two South Indian nation-states: Tamil Nadu and Kerala [72, 73, 82]. For each, we selected objects that are culturally significant to the community but not commonly known outside of the culture, placing them in the long tail. Specifically, we selected a *braille notetaker*, a portable electronic device used daily by many people who are blind; and a *Mridangam*, a percussion instrument from Tamil Nadu that is central to South Indian classical music. See Figure 1.

During online synchronous workshops, community members were shown 5 AI-generated images of each object and asked to (1) *rate* whether the image was a correct depiction of the object on a scale of 1 to 3; and (2) *rank* the images in order from most to least preferred. Images were shown one at a time, and participants were invited to reflect on the motivations for their decisions for each image. We conducted 14 interviews with blind and low vision community members (using manually generated alt text for the images we presented), and ran focus groups with 17 total residents of Tamil Nadu or Kerala. We provide our complete study protocol for both communities in Appendix B.

4 Methodological Challenges of Validating Measures at the Long Tail

In attempting to correlate image-text alignment metrics for the generated images with the human preference data from our two case studies, following standard practices, we encountered two significant challenges.

Here we discuss each challenge and its implications for validating metrics at the long tail.

4.1 Challenge: Scaling community annotations

Annotators require community knowledge. Blind community members consistently drew upon their embodied knowledge of how a braille notetaker is used to assess if an AI depiction was plausible. For example, a braillist would know that a tactile braille display must be a single horizontal line of cells positioned below the keys. Otherwise, a user may accidentally "knock the keys" when trying to read with their hands. This *embodied knowledge* helped community members prioritize functional characteristics, while accounting for legitimate variation in the visual characteristics of color, shape, and size.

In ranking tasks specifically, annotators had to make value judgments about which errors were "most wrong." To achieve this, blind community members often drew upon their situated experiences as disabled individuals in the world. We learned that depictions of braille on an electronic screen, for example, are substantially less preferred because electronic screens are inaccessible to blind users. Thus, annotators' rankings were influenced not only by their knowledge about each object, but by the *historical context* of exclusion that they and their community had experienced.

When working with residents of Tamil Nadu, we found that annotators similarly drew upon both embodied knowledge and historical context for the Indian classical music percussion instrument, the Mridangam. For example, community members noted that the Mridangam is always played horizontally and typically placed on one's lap during a performance. Images that show the instrument standing upright or resting on the floor (Figure 1) were dismissed as implausible, as the instrument's shape does not support such positioning. As one community member said, "*It can't be placed on floor like this. It will roll due to its steeper curve and imbalance between the two ends.*" This understanding extended to how people were depicted with the artifact. Images where individuals held the instrument incorrectly were seen as misrepresenting cultural norms, with community members noting that such portrayals gave the impression of a different instrument, such as a kick drum, rather than a Mridangam.

Common practices assume that data annotation, including rating and ranking tasks, can be completed by any cognizant annotator. Yet, these two examples challenge this assumption, highlighting how embodied knowledge and historical context are essential for accurate annotation — particularly in cases where that knowledge is unfamiliar to mainstream groups. This finding augments the growing body of literature (Section 2.3) that calls for situated knowledge and expertise in AI evaluation tasks.

Annotations place burdens on communities. Validating a metric using human preference data requires a minimum sample size to ensure statistical reliability [36, 41, 64]. In conventional crowdsourced annotation pipelines, this typically large workload is distributed across a pool of annotators, minimizing the burden on any individual [64]. However, annotators with community knowledge, which we've argued above is necessary, might be a much smaller pool [56, 65]. As reliability of conclusions depends on both the number of images annotated and the number of annotators per image, this could raise the burden substantially for any one individual.

Consider an example estimate for the number of annotations that would be required to validate a metric using a dataset of prompts (e.g., depictions of cultural artifacts) curated by a community (Figure 3, Left), using Spearman's rank correlation. Suppose we aim to confirm that a computed $\rho = 0.7$ falls within a ± 0.05 margin of error at a 95% confidence interval. Using a standard error approximation (Appendix C), this would require at least 402 samples (i.e., generated images with human preference labels). To account for inter-annotator disagreement, each sample typically receives labels from three annotators, with the majority (or average) vote used as the final label [64]. This would amount to a total requirement of 1206 individual annotations.

While collecting this number of annotations might be feasible, the time and effort required to do so will vary significantly across communities, especially for marginalized communities who may require specific accommodations. In our study with the blind and low vision community, we observed that most community members spent a minimum of 5 minutes reviewing and providing annotations (both rankings and ratings) for each generated image. Assuming that we would need a minimum of 1206 annotations, this would amount to a total of around 100 hours of community member time annotating images of relevant concepts.

Although the blind and low vision community has unique accessibility needs (e.g., requiring images to be described in words [16, 40]), this level of effort is not unique to this community. Other groups may require language translation support [5] or technical support [31] to facilitate participation. Community annotators may not be familiar with AI and may need a tutorial to provide appropriate annotations [68, 98]. These time estimates also do not include the additional effort required for recruitment, on-boarding, and training. All of these provisions are essential to ethical and effective engagement with marginalized communities [10, 14, 31, 68, 98].

This example is likely a significant underestimate of the time and effort that would be needed to make this work in practice. In real-world scenarios, the correlation ρ between metric scores and human preferences is often lower than the above idealized example [38, 51]. This would require more samples to achieve the same level of statistical significance, given that ρ and n are inversely and quadratically related. Moreover, when comparing models with similar performance, the evaluation metric must be precise enough to detect subtle differences — often necessitating a higher bar for metric validation. For example, recent work [36] has shown that for models whose outputs are stochastic, between 25,000 and 50,000 total ratings were needed to achieve statistical significance. This would be untenable for communities from whom even the conservative estimates are already substantial. In such cases, metric validation would become impossible.

4.2 Challenge: Navigating "shades of bad"

Existing metric validation approaches that use human preferences assume that the annotated images reflect a meaningful range of representations, so that annotators can distinguish between good and bad depictions [33, 64]. However, current state-of-the-art T2I models often fail to generate accurate depictions of tail concepts. In both of our case studies, we were unable to generate even semi-accurate depictions of braille notetakers or the Mridangam, despite attempting interventions (e.g., using prompt engineering techniques [66, 71]) detailed in Appendix B. Instead of displaying a meaningful range of good and bad depictions, all the generated images had at least one substantial error: a phenomenon we term "shades of bad."

Only having bad representations available to show annotators poses fundamental challenges in using collected rating and ranking data. For rating tasks, if annotators consistently provide low ratings, the data will lack the meaningful variance needed to validate metrics. In our case studies, members of both communities consistently assigned low scores to generated images (Table 1). All generated images of both objects had average ratings (on a 3-point scale) between a 1 ("The image is totally unlike the object") and a 2 ("The image is partially correct"). None of the generated images of either object were rated as accurate depictions of the object. While a rank correlation can be calculated from heavily imbalanced ratings, because there are no highly rated images, the correlation would fail to capture if a correct depiction would be scored higher. Thus, while we may be able to validate if a metric can assign low scores to poor depictions, we cannot confirm that it would preference correct depictions, a necessary behavior to be able to steer models towards improved representations.

For ranking tasks, shades of bad poses even more foundational challenges for metric validation. If all generations are poor, ordering them becomes a task of "which error is worse." While blind community members were quick to assess that these images were wrong, they often struggled to make what often felt like arbitrary judgment calls. This uncertainty was reflected in very high variance across annotators' *rankings* (detailed in Appendix D), despite high consistency in their *ratings*.

Shades of bad poses fundamental challenges to the utility of data collected from all available rating and ranking tasks, resulting in imbalanced ratings and arbitrary rankings. As a result, while we can use the collected data to validate if our metrics can assign low scores to poor depictions, we still do not know whether our metrics can tell us if we're making progress towards improved depictions — a prerequisite to address critical errors. Moreover, community members who were repeatedly exposed to incorrect or even offensive depictions found the annotation task to be difficult and demoralizing. As a result, the collected data leaves us in the same place we started: without validated metrics.

5 Call to Action: Rethinking T2I Measurement Practices

State-of-the-art T2I evaluation metrics often rely on secondary models that systematically fail to recognize tail concepts [60, 85], underscoring the need for robust validation. Yet, our analysis reveals how existing validation methodologies that rely on crowdsourced rating or ranking annotations break down when applied to tail concepts. Capturing preferences of generated images of tail concepts requires annotators with cultural knowledge, which standard crowdsourcing pipelines are not designed to support [43]. Recruiting such annotators at scale is resource-intensive and in some cases, infeasible. Furthermore, when the available generations are all misrepresentations ("shades of bad"), annotations offer limited value beyond confirming that a metric penalizes errors. Our work has confirmed this to be the case for two independent marginalized communities we engaged.

While these challenges may not appear in all low-data contexts, they are likely to resonate with practitioners engaging with populations currently underserved by frontier models. As such, the challenges we've surfaced offer a useful lens for practitioners to critically examine their own approaches to measure validation. Before collecting data, practitioners should consider both the domain-specific knowledge required and number of annotations needed, as well as whether available T2I model outputs provide a meaningful range of good and bad representations to annotate. These challenges may manifest differently across contexts – e.g., some communities may be easier to engage than others – but without reasonable depictions to annotate, annotations offer limited utility in validating metrics.

In this section, we discuss directions for future research in light of the methodological challenges we have surfaced. Our first set of research directions seeks to incrementally address these challenges by modifying existing evaluation practices. Our second set of research directions explores fundamentally different ways to think about T2I evaluation by centering community expertise.

5.1 Innovating within our existing measurement practices

We first propose directions for methodological innovations that may lead to improvements within existing measurement processes.

Improving existing T2I metrics. Future research can further investigate how biases in large pretrained models might propagate to downstream evaluation metrics, and explore how metrics might be modified to address them. For instance, the CLIPScore metric works by calculating the cosine similarity score between text and image embeddings [34]. Future research can experiment with proposing modifications to the metric targeted towards improving performance for low data concepts, similarly to [60, 103]. To name a few possibilities, one alternative metric could explore applying bias mitigation techniques to modify the CLIP text embedding used to score output images [7], e.g., so that the text embedding does not preference images that depict "notetaking" using paper. Another alternative metric could use reference photos (e.g., of a braille notetaker) and their CLIP *image* embeddings (in place of biased text embeddings) to score generated images [99].

Scaling up annotations. Future research can explore methods to reduce the annotation load for community members. Instead of asking the community to provide hundreds of annotations, researchers can design alternative workflows. For example, to increase efficiency, researchers and community members might first sort AI outputs into clusters that share common characteristics and then ask community members to annotate each cluster with shared judgments (e.g., "depictions of writing with ink on paper are *all* inaccurate depictions of a braille notetaker"). Future work can draw from HCI methods to explore other collaborative workflows that combine in-group and out-of-group annotators to leverage community expertise where it is most needed.

Exploring the affordances of real photographs for metric validation. To address "shades of bad," practitioners can explore alternative methods to confirm if metrics preference correct depictions of tail concepts. One sanity check could involve designing controlled experiments to validate that a metric assigns higher scores to actual photographs of the concept (e.g., a braille notetaker), relative to a comparison group (e.g., photos of related objects such as a paper notebook), following Massiceti et al. [60]. However, using real photographs as a proxy has several limitations. First, the distribution shift between photographs versus AI images and its effect on metric scores is unclear [15]. Second, while a real photograph either is or isn't a braille notetaker, AI images exist in the world of the imaginary — producing errors and depictions that do not exist in the real world [16]. Thus, further validation is still needed to understand how an ideal metric would score AI depictions.

5.2 Re-imagining measurement for the long tail

Longer term, we advocate for exploring fundamentally different ways to think about T2I evaluation, drawing on emerging calls for AI evaluation to learn from the social sciences [73, 87, 101] and acknowledging that evaluation is iterative, with room for community engagement at multiple stages.

Drawing on the framework of Adcock and Collier [1], Wallach et al. [87] propose that approaches to AI evaluation should clearly distinguish between the process of *systematization* — taking a broad concept like "image-prompt alignment" or "appropriateness" and narrowing it down into an explicit definition — from the process of *operationalization* — specifying the procedures or metrics that will be used to obtain measurements of the concept. This allows the separation of conceptual debates (for instance, does our definition of "appropriateness" reflect what we want it to reflect?) from operational debates (does the metric we have chosen yield a valid measurement?). In addition to bringing more rigor to the measurement process, Wallach et al. [87] argue that this approach can open up new opportunities for community participation in the measurement process by including people with different expertise and experiences in conceptual debates.

In contrast, status quo approaches to evaluating T2I models bypass systematization, moving straight from a background concept like "image-prompt alignment" to metrics that are typically chosen based on convenience and availability with little thought to exactly what these metrics do or do not capture. Any validation of these metrics occurs late in the process. Furthermore, because validation is at the level of the metric only, it tends to be limited to quantitative approaches, like computing correlation with crowdsourced ratings, as described in Section 2.2.

To move beyond crowdsourced annotations as a proxy for "ground truth," we encourage future work to explore alternative ways of eliciting (and translating) community knowledge at earlier stages of the evaluation process. For example, researchers can incorporate community participation at the systematization stage to define what accurate depictions of tail concepts should (and should not) include. Inviting community members to participate in the iterative process of constructing and critiquing a shared systematization may allow us to make better use of their unique expertise to inform what metrics are designed in the first place.

While prior work on metric validation has relied on structured annotation formats such as ratings or rankings [64], our community engagements surfaced rich forms of community expertise that closed-form annotation tasks fail to capture. Thus, even when community members struggled to rank inaccurate depictions of a braille notetaker, they still held a clear and shared understanding of what it should look like, even if no such images were available to annotate. Such knowledge was not reflected in the collected rating and ranking data, especially when all the images were bad. We urge researchers to explore methods to incorporate such "thick descriptions" [73] of generated images to create and validate new measures, rather than treating community members akin to anonymous crowdworkers.

6 Alternative Views

In this section, we respond to alternative views to our work's focus on improving measurement practices for the long tail.

Q: Why should we care about evaluation at the long tail when the capabilities of frontier models keep improving? Why don't we just wait until capabilities get better? Frontier models are improving, but this does not invalidate the reality that they underperform for many marginalized communities today. Evaluation metrics that can reliably distinguish offensive from accurate outputs can inform safety mitigations [90, 95] that can impact the millions of AI images generated each day [25, 86]. Furthermore, achieving improved representations of tail concepts still remains an open challenge. For instance, Massiceti et al. [60] showed that many assistive technologies appear fewer than 20 times in large pretraining datasets, well below the threshold needed for models to learn them effectively [85]. As a result, disparities between closed and open-source models may widen due to the scarcity of publicly available data [53, 54, 80]. Thus, effective measurement remains essential not only to mitigate harms in today's deployed systems, but also to enable meaningful scientific progress at making tomorrow's models work better, for everyone.

Q: Won't "Shades of Bad" be addressed by the release of new and improved frontier models? While model capabilities may improve for some concepts, capabilities may be slower to improve for others. Indeed, the concepts where measurement has the potential to be most impactful are the ones that the frontier models continue to get wrong. Additionally, the existence of inherently low-data scenarios, such as personalized generation tasks [12, 23, 83], ensures that long-tail problems will persist. The need for valid evaluation at the frontier remains.

Q: Do we really need community members to participate as annotators? Why can't I train people outside of the community? While knowledgeable outsiders may be able to serve as community "proxies" in some contexts, outsiders often lack the embodied experience and historical understanding required to make informed judgements (Section 4.1). Research shows that without this foundation, out-group annotators may project misguided assumptions when making judgments, replicating existing patterns of oppression in media [25, 30, 58]. At stake is not only annotation quality, but also the question of *who* gets to define how a community is represented: a central concern in participatory AI [19, 73, 84, 96]. Recognizing and centering community members' expertise is a key first step towards developing more inclusive AI systems.

Q: I don't encounter these challenges in my context. Can't I just follow the same blueprint as everyone else? We anticipate that the challenges surfaced in Section 4 may not affect every community or tail concept. When such obstacles are absent, applying conventional validation techniques using rating and ranking data is appropriate. Even in such settings, we believe that practitioners may benefit from exploring alternative metric validation methods such as those we have proposed.

7 Acknowledgments

We thank our study participants whose time and expertise made our research possible. We thank members of the MSR Teachable AI Experiences Team for their support throughout this work. We thank attendees of the CMU Fairness, Ethics, Accountability, and Transparency reading group for providing feedback on our manuscript.

References

[1] R. Adcock and D. Collier. Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3):529–546, 2001.

[2] L. Aroyo and C. Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, Mar. 2015. doi: 10.1609/aimag.v36i1.2564. URL `https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2564`.

[3] L. Aroyo, L. Dixon, N. Thain, O. Redfield, and R. Rosen. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1100–1105, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366755. doi: 10.1145/3308560.3317083. URL `https://doi.org/10.1145/3308560.3317083`.

[4] L. Aroyo, A. Taylor, M. Díaz, C. Homan, A. Parrish, G. Serapio-García, V. Prabhakaran, and D. Wang. Dices dataset: Diversity in conversational ai evaluation for safety. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 53330–53342. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/a74b697bce4cac6c91896372abaa8863-Paper-Datasets_and_Benchmarks.pdf`.

[5] K. Bali, M. Choudhury, and V. Seshadri. Ellora: Enabling low resource languages with technology. *Proceedings of the 1st International Conference on Language Technologies for All*, 2019. URL `https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.41.pdf`.

[6] C. L. Bennett, E. Brady, and S. M. Branham. Interdependence as a frame for assistive technology research and design. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '18, page 161–173, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356503. doi: 10.1145/3234695.3236348. URL `https://doi.org/10.1145/3234695.3236348`.

[7] H. Berg, S. M. Hall, Y. Bhalgat, W. Yang, H. R. Kirk, A. Shtedritski, and M. Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. *arXiv preprint arXiv:2203.11933*, 2022.

[8] S. Bergman, N. Marchal, J. Mellor, S. Mohamed, I. Gabriel, and W. Isaac. Stela: a community-centred approach to norm elicitation for ai alignment. *Scientific Reports*, 14, 03 2024. doi: 10.1038/s41598-024-56648-4.

[9] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1493–1504, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594095. URL `https://doi.org/10.1145/3593013.3594095`.

[10] A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, and S. Mohamed. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394772. doi: 10.1145/3551624.3555290. URL `https://doi.org/10.1145/3551624.3555290`.

[11] D. G. Bonett. Meta-analytic interval estimation for bivariate correlations. *Psychological Methods*, 13(3):173, 2008.

[12] A. Bose, Z. Xiong, Y. Chi, S. S. Du, L. Xiao, and M. Fazel. Lore: Personalizing llms via low-rank reward modeling, 2025. URL `https://arxiv.org/abs/2504.14439`.

[13] J. Chien and D. Danks. Beyond behaviorist representational harms: A plan for measurement and mitigation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 933–946, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658946. URL `https://doi.org/10.1145/3630106.3658946`.

[14] N. Cooper, T. Horne, G. R. Hayes, C. Heldreth, M. Lahav, J. Holbrook, and L. Wilcox. A systematic review and thematic analysis of community-collaborative approaches to computing research. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517716. URL `https://doi.org/10.1145/3491102.3517716`.

[15] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva. Raising the bar of ai-generated image detection with clip. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4356–4366, 2024. doi: 10.1109/CVPRW63382.2024.00439.

[16] M. Das, A. J. Fiannaca, M. R. Morris, S. K. Kane, and C. L. Bennett. From provenance to aberrations: Image creator and screen reader user perspectives on alt text for ai-generated images. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642325. URL `https://doi.org/10.1145/3613904.3642325`.

[17] S. Datta, A. Ku, D. Ramachandran, and P. Anderson. Prompt expansion for adaptive text-to-image generation. *arXiv preprint arXiv:2312.16720*, 2023.

[18] S. Dehdashtian, G. Sreekumar, and V. N. Boddeti. Oasis uncovers: High-quality t2i models, same old stereotypes, 2025. URL `https://arxiv.org/abs/2501.00962`.

[19] F. Delgado, S. Yang, M. Madaio, and Q. Yang. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400703812. doi: 10.1145/3617694.3623261. URL `https://doi.org/10.1145/3617694.3623261`.

[20] W. H. Deng, M. Yurrita, M. Díaz, J. Suh, N. Judd, L. Groves, H. Shen, M. Eslami, and K. Holstein. Responsible crowdsourcing for responsible generative ai: Engaging crowds in ai auditing and evaluation. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 12(1):148–150, Oct. 2024. doi: 10.1609/hcomp.v12i1.31609. URL `https://ojs.aaai.org/index.php/HCOMP/article/view/31609`.

[21] A. Deviyani and F. Diaz. Contextual metric meta-evaluation by measuring local metric accuracy. *arXiv preprint arXiv:2503.19828*, 2025.

[22] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, and R. Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12606–12633. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/esser24a.html`.

[23] V. Gallego. Personalizing text-to-image generation via aesthetic gradients. *arXiv preprint arXiv:2209.12330*, 2022.

[24] S. Gautam, P. N. Venkit, and S. Ghosh. From melting pots to misrepresentations: Exploring harms in Generative AI. *arXiv preprint arXiv:2403.10776*, 2024.

[25] T. Gillespie. Generative AI and the Politics of Visibility. *Big Data & Society*, 11(2): 20539517241252131, June 2024. doi: 10.1177/20539517241252131.

[26] T. Gnambs. A brief note on the standard error of the pearson correlation. *Collabra: Psychology*, 9(1):87615, 2023.

[27] N. Goyal, I. D. Kivlichan, R. Rosen, and L. Vasserman. Is your toxicity my toxicity? Exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28, 2022.

[28] L. Guerdan, S. Barocas, K. Holstein, H. Wallach, Z. S. Wu, and A. Chouldechova. Validating LLM-as-a-Judge Systems in the Absence of Gold Labels. *arXiv preprint arXiv:2503.05965*, 2025.

[29] R. Hada, V. Gumma, A. de Wynter, H. Diddee, M. Ahmed, M. Choudhury, K. Bali, and S. Sitaram. Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation? *arXiv preprint arXiv:2309.07462*, 2024.

[30] M. Hall, S. J. Bell, C. Ross, A. Williams, M. Drozdzal, and A. R. Soriano. Towards geographic inclusion in the evaluation of text-to-image models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 585–601, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658927. URL `https://doi.org/10.1145/3630106.3658927`.

[31] S. M. Hall, S. Dalal, R. Sefala, F. Yuehgoh, A. Alaagib, I. Hamzaoui, S. Ishida, J. Magomere, L. Crais, A. Salama, and T. Afonja. The Human Labour of Data Work: Capturing Cultural Diversity through World Wide Dishes. *arXiv preprint arXiv:2502.05961*, 2025.

[32] A. Hardy, A. Reuel, K. Jafari Meimandi, L. Soder, A. Griffith, D. M. Asmar, S. Koyejo, M. S. Bernstein, and M. J. Kochenderfer. More than Marketing? On the Information Value of AI Benchmarks for Practitioners. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 1032–1047, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713064. doi: 10.1145/3708359.3712152. URL `https://doi.org/10.1145/3708359.3712152`.

[33] S. Hartwig, D. Engel, L. Sick, H. Kniesel, T. Payer, P. Poonam, M. Glöckler, A. Bäuerle, and T. Ropinski. A survey on quality metrics for text-to-image generation, 2025. URL `https://arxiv.org/abs/2403.11821`.

[34] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[35] C. Homan, G. Serapio-Garcia, L. Aroyo, M. Diaz, A. Parrish, V. Prabhakaran, A. Taylor, and D. Wang. Intersectionality in AI safety: Using multilevel models to understand diverse perceptions of safety in conversational AI. In G. Abercrombie, V. Basile, D. Bernadi, S. Dudy, S. Frenda, L. Havens, and S. Tonelli, editors, *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 131–141, Torino, Italia, May 2024. ELRA and ICCL. URL `https://aclanthology.org/2024.nlperspectives-1.15/`.

[36] C. M. Homan, S. Wein, C. Welty, and L. Aroyo. How Many Raters Do You Need? Power Analysis for Foundation Models. In *NeurIPS 2023 Workshop on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models"*, 2023.

[37] R. Hong, W. Agnew, T. Kohno, and J. Morgenstern. Who's in and who's out? A case study of multimodal CLIP-filtering in DataComp. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400712227. doi: 10.1145/3689904.3694702. URL `https://doi.org/10.1145/3689904.3694702`.

[38] Y. Hu, B. Liu, J. Kasai, Y. Wang, M. Ostendorf, R. Krishna, and N. A. Smith. TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023.

[39] K. Huang, C. Duan, K. Sun, E. Xie, Z. Li, and X. Liu. T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[40] M. Huh, Y.-H. Peng, and A. Pavel. Genassist: Making image generation accessible. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606735. URL `https://doi.org/10.1145/3586183.3606735`.

[41] O. Inel, T. Draws, and L. Aroyo. Collect, measure, repeat: Reliability factors for responsible ai data collection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 11(1):51–64, Nov. 2023. doi: 10.1609/hcomp.v11i1.27547. URL `https://ojs.aaai.org/index.php/HCOMP/article/view/27547`.

[42] A. Jha, V. Prabhakaran, R. Denton, S. Laszlo, S. Dave, R. Qadri, C. K. Reddy, and S. Dev. ViSAGe: A global-scale analysis of visual stereotypes in text-to-image generation. *arXiv preprint arXiv:2401.06310*, 2024.

[43] S. Kapania, A. S. Taylor, and D. Wang. A hunt for the snark: Annotator diversity in data practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3580645. URL `https://doi.org/10.1145/3544548.3580645`.

[44] S. Kapania, S. Ballard, A. Kessler, and J. W. Vaughan. Examining the expanding role of synthetic data throughout the ai development pipeline. In *Proceedings of the 8th ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2025.

[45] J. Katzman, A. Wang, M. Scheuerman, S. L. Blodgett, K. Laird, H. Wallach, and S. Barocas. Taxonomizing and Measuring Representational Harms: A Look at Image Tagging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14277–14285, 2023.

[46] S. Kingsley, J. Zhi, W. H. Deng, J. Lee, S. Zhang, M. Eslami, K. Holstein, J. I. Hong, T. Li, and H. Shen. Investigating what factors influence users' rating of harmful algorithmic bias and discrimination. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 12(1):75–85, Oct. 2024. doi: 10.1609/hcomp.v12i1.31602. URL `https://ojs.aaai.org/index.php/HCOMP/article/view/31602`.

[47] H. R. Kirk, A. Whitefield, P. Rottger, A. M. Bean, K. Margatina, R. Mosquera-Gomez, J. Ciro, M. Bartolo, A. Williams, H. He, et al. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37: 105236–105344, 2024.

[48] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

[49] E. Kreiss, C. Bennett, S. Hooshmand, E. Zelikman, M. R. Morris, and C. Potts. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. *arXiv preprint arXiv:2205.10646*, 2022.

[50] M. Krumdick, C. Lovering, V. Reddy, S. Ebner, and C. Tanner. No Free Labels: Limitations of LLM-as-a-Judge Without Human Grounding. *arXiv preprint arXiv:2503.05061*, 2025.

[51] T. Lee, M. Yasunaga, C. Meng, Y. Mai, J. S. Park, A. Gupta, Y. Zhang, D. Narayanan, H. Teufel, M. Bellagente, et al. Holistic Evaluation of Text-To-Image Models. *Advances in Neural Information Processing Systems*, 36:69981–70011, 2023.

[52] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[53] S. Longpre, R. Mahari, A. Lee, C. Lund, H. Oderinwale, W. Brannon, N. Saxena, N. Obeng-Marnu, T. South, C. Hunter, K. Klyman, C. Klamm, H. Schoelkopf, N. Singh, M. Cherep, A. M. Anis, A. Dinh, C. Chitongo, D. Yin, D. Sileo, D. Mataciunas, D. Misra, E. Alghamdi,

E. Shippole, J. Zhang, J. Materzynska, K. Qian, K. Tiwary, L. Miranda, M. Dey, M. Liang, M. Hamdy, N. Muennighoff, S. Ye, S. Kim, S. Mohanty, V. Gupta, V. Sharma, V. M. Chien, X. Zhou, Y. Li, C. Xiong, L. Villa, S. Biderman, H. Li, D. Ippolito, S. Hooker, J. Kabbara, and S. Pentland. Consent in Crisis: The Rapid Decline of the AI Data Commons. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 108042–108087. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/c3738949a80306cc48a8ea8ba0560f9d-Paper-Datasets_and_Benchmarks_Track.pdf`.

[54] S. Longpre, N. Singh, M. Cherep, K. Tiwary, J. Materzynska, W. Brannon, R. Mahari, N. Obeng-Marnu, M. Dey, M. Hamdy, et al. Bridging the Data Provenance Gap Across Text, Speech and Video. *arXiv preprint arXiv:2412.17847*, 2024.

[55] Y. Lu, X. Yang, X. Li, X. E. Wang, and W. Y. Wang. LLMScore: Unveiling the Power of Large Language Models in Text-to-Image Synthesis Evaluation. *Advances in Neural Information Processing Systems*, 36:23075–23093, 2023.

[56] K. Mack, E. McDonnell, D. Jain, L. Lu Wang, J. E. Froehlich, and L. Findlater. What do we mean by "accessibility research"? a literature survey of accessibility papers in chi and assets from 1994 to 2019. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445412. URL `https://doi.org/10.1145/3411764.3445412`.

[57] K. A. Mack, R. Qadri, R. Denton, S. K. Kane, and C. L. Bennett. "They only care to show us the wheelchair": disability representation in text-to-image AI models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642166. URL `https://doi.org/10.1145/3613904.3642166`.

[58] S. S. A. Magid, W. Pan, S. Warchol, G. Guo, J. Kim, M. Rahman, and H. Pfister. Is what you ask for what you get? investigating concept associations in text-to-image models, 2025. URL `https://arxiv.org/abs/2410.04634`.

[59] J. Magomere, S. Ishida, T. Afonja, A. Salama, D. Kochin, F. Yuehgoh, I. Hamzaoui, R. Sefala, A. Alaagib, S. Dalal, B. Marchegiani, E. Semenova, L. Crais, and S. M. Hall. The World Wide Recipe: A community-centred framework for fine-grained data collection and regional bias operationalisation. *arXiv preprint arXiv:2406.09496*, 2025.

[60] D. Massiceti, C. Longden, A. Slowik, S. Wills, M. Grayson, and C. Morrison. Explaining CLIP's Performance Disparities on Data from Blind/Low Vision Users. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12172–12182, June 2024.

[61] N. J. Mim, D. Nandi, S. S. Khan, A. Dey, and S. I. Ahmed. In-Between Visuals and Visible: The Impacts of Text-to-Image Generative AI Tools on Digital Image-making Practices in the Global South. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3641951. URL `https://doi.org/10.1145/3613904.3641951`.

[62] A. Muehlbradt and S. K. Kane. What's in an ALT Tag? Exploring Caption Content Priorities through Collaborative Captioning. *ACM Trans. Access. Comput.*, 15(1), Mar. 2022. ISSN 1936-7228. doi: 10.1145/3507659. URL `https://doi.org/10.1145/3507659`.

[63] T. Nguyen, M. Wallingford, S. Santy, W.-C. Ma, S. Oh, L. Schmidt, P. W. Koh, and R. Krishna. Multilingual diversity improves vision-language representations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 91430–91459. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/a6678e2be4ce7aef9d2192e03cd586b7-Paper-Conference.pdf`.

[64] M. Otani, R. Togashi, Y. Sawai, R. Ishigami, Y. Nakashima, E. Rahtu, J. Heikkilä, and S. Satoh. Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14277–14286, 2023.

[65] L. A. Palinkas, S. M. Horwitz, C. A. Green, J. P. Wisdom, N. Duan, and K. Hoagwood. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health and Mental Health Services Research*, 42(5):533–544, September 2015. doi: 10.1007/s10488-013-0528-y. URL `https://pubmed.ncbi.nlm.nih.gov/24193818/`.

[66] S. Parashar, Z. Lin, T. Liu, X. Dong, Y. Li, D. Ramanan, J. Caverlee, and S. Kong. The Neglected Tails in Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12988–12997, 2024.

[67] A. Parrish, S. Hao, S. Laszlo, and L. Aroyo. "Is a picture of a bird a bird?" A mixed-methods approach to understanding diverse human perspectives and ambiguity in machine vision models. In G. Abercrombie, V. Basile, D. Bernadi, S. Dudy, S. Frenda, L. Havens, and S. Tonelli, editors, *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 1–18, Torino, Italia, May 2024. ELRA and ICCL. URL `https://aclanthology.org/2024.nlperspectives-1.1/`.

[68] Partnership on AI. Guidance for inclusive ai: Practicing participatory engagement, 2025. URL `https://partnershiponai.org/guidance-for-inclusive-ai/`.

[69] S. Pawar, J. Park, J. Jin, A. Arora, J. Myung, S. Yadav, F. G. Haznitrama, I. Song, A. Oh, and I. Augenstein. Survey of Cultural Awareness in Language Models: Text and Beyond. *Computational Linguistics*, pages 1–96, 2025.

[70] V. Prabhakaran, C. Homan, L. Aroyo, A. M. Davani, A. Parrish, A. Taylor, M. Díaz, D. Wang, and G. Serapio-García. GRASP: A Disagreement Analysis Framework to Assess Group Associations in Perspectives. *arXiv preprint arXiv:2311.05074*, 2024.

[71] S. Pratt, I. Covert, R. Liu, and A. Farhadi. What Does a Platypus Look Like? Generating Customized Prompts for Zero-Shot Image Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023.

[72] R. Qadri, R. Shelby, C. L. Bennett, and R. Denton. AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 506–517, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594016. URL `https://doi.org/10.1145/3593013.3594016`.

[73] R. Qadri, M. Diaz, D. Wang, and M. Madaio. The Case for "Thick Evaluations" of Cultural Representation in AI. *arXiv preprint arXiv:2503.19075*, 2025.

[74] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[75] D. Raji, E. Denton, E. M. Bender, A. Hanna, and A. Paullada. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf`.

[76] C. Rastogi, T. H. Teh, P. Mishra, R. Patel, Z. Ashwood, A. M. Davani, M. Diaz, M. Paganini, A. Parrish, D. Wang, V. Prabhakaran, L. Aroyo, and V. Rieser. Insights on disagreement patterns in multimodal safety perception across diverse rater groups, 2024. URL `https://arxiv.org/abs/2410.17032`.

[77] C. Ross, M. Hall, A. R. Soriano, and A. Williams. What makes a good metric? Evaluating automatic metrics for text-to-image consistency. *arXiv preprint arXiv:2412.13989*, 2024.

[78] Royal National Institute of Blind People. Braille displays and readers, 2025. URL `https://www.rnib.org.uk/living-with-sight-loss/assistive-aids-and-technology/reading-and-writing/an-rnib-guide-to-braille-displays-for-blind-and-partially-sighted-people/`.

[79] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *Advances in neural information processing systems*, 35: 36479–36494, 2022.

[80] M. Saxon, F. Jahara, M. Khoshnoodi, Y. Lu, A. Sharma, and W. Y. Wang. Who Evaluates the Evaluations? Objectively Scoring Text-to-Image Prompt Coherence Metrics with T2IScoreScore (TS2). In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 85630–85657. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/9b9cfd5428153ccfbd4ba34b7e007305-Paper-Conference.pdf`.

[81] N. K. Senthilkumar, A. Ahmad, M. Andreetto, V. Prabhakaran, U. Prabhu, A. B. Dieng, P. Bhattacharyya, and S. Dave. Beyond Aesthetics: Cultural Competence in Text-to-Image Models. *Advances in Neural Information Processing Systems*, 37:13716–13747, 2024.

[82] A. Seth, S. Ahuja, K. Bali, and S. Sitaram. DOSA: A Dataset of Social Artifacts from Different Indian Geographical Subcultures. *arXiv preprint arXiv:2403.14651*, 2024.

[83] A. Singh, S. Hsu, K. Hsu, E. Mitchell, S. Ermon, T. Hashimoto, A. Sharma, and C. Finn. FSPO: Few-Shot Preference Optimization of Synthetic Preference Data in LLMs Elicits Effective Personalization to Real Users. *arXiv preprint arXiv:2502.19312*, 2025.

[84] H. Suresh, E. Tseng, M. Young, M. Gray, E. Pierson, and K. Levy. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1609–1621, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658992. URL `https://doi.org/10.1145/3630106.3658992`.

[85] V. Udandarao, A. Prabhu, A. Ghosh, Y. Sharma, P. H. Torr, A. Bibi, S. Albanie, and M. Bethge. No "Zero-Shot" Without Exponential Data: Pretraining Concept Frequency Determines Multimodal Model Performance. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 61735–61792. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/715b78ccfb6f4cada5528ac9b5278def-Paper-Conference.pdf`.

[86] A. Valyaeva. People are creating an average of 34 million images per day, 2024. URL `https://journal.everypixel.com/ai-image-statistics/`.

[87] H. Wallach, M. Desai, A. F. Cooper, A. Wang, C. Atalla, S. Barocas, S. L. Blodgett, A. Chouldechova, E. Corvi, P. A. Dow, J. Garcia-Gathright, A. Olteanu, N. Pangakis, S. Reed, E. Sheng, D. Vann, J. W. Vaughan, M. Vogel, H. Washington, and A. Z. Jacobs. Position: Evaluating Generative AI Systems is a Social Science Measurement Challenge. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025. URL `https://arxiv.org/abs/2502.00561`.

[88] D. Wang, M. Díaz, A. Parrish, L. Aroyo, C. Homan, G. Serapio-García, V. Prabhakaran, and A. Taylor. All that agrees is not gold: Evaluating ground truth labels and dialogue content for safety, 2023.

[89] J. Warren, G. M. Weiss, F. Martinez, A. Guo, and Y. Zhao. Decoding fatphobia: Examining anti-fat and pro-thin bias in AI-generated images. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4724–4736, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. URL `https://aclanthology.org/2025.findings-naacl.266/`.

[90] L. Weidinger, M. Rauh, N. Marchal, A. Manzini, L. A. Hendricks, J. Mateos-Garcia, S. Bergman, J. Kay, C. Griffin, B. Bariach, et al. Sociotechnical Safety Evaluation of Generative AI Systems. *arXiv preprint arXiv:2310.11986*, 2023.

[91] L. Weidinger, I. D. Raji, H. Wallach, M. Mitchell, A. Wang, O. Salaudeen, R. Bommasani, D. Ganguli, S. Koyejo, and W. Isaac. Toward an Evaluation Science for Generative AI Systems. *arXiv preprint arXiv:2503.05336*, 2025.

[92] L. Xiao, M. Bandukda, K. Angerbauer, W. Lin, T. Bhatnagar, M. Sedlmair, and C. Holloway. A Systematic Review of Ability-diverse Collaboration through Ability-based Lens in HCI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3641930. URL `https://doi.org/10.1145/3613904.3641930`.

[93] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023. URL `https://arxiv.org/abs/2304.05977`.

[94] J. Ye, Y. Wang, Y. Huang, D. Chen, Q. Zhang, N. Moniz, T. Gao, W. Geyer, C. Huang, P.-Y. Chen, N. V. Chawla, and X. Zhang. Justice or prejudice? quantifying biases in llm-as-a-judge, 2024. URL `https://arxiv.org/abs/2410.02736`.

[95] J. Yoon, S. Yu, V. Patil, H. Yao, and M. Bansal. SAFREE: Training-Free and Adaptive Guard for Safe Text-to-Image And Video Generation. *arXiv preprint arXiv:2410.12761*, 2025.

[96] M. Young, U. Ehsan, R. Singh, E. Tafesse, M. Gilman, C. Harrington, and J. Metcalf. Participation versus scale: Tensions in the practical demands on participatory ai. *First Monday*, 29(4), Apr. 2024. doi: 10.5210/fm.v29i4.13642. URL `https://firstmonday.org/ojs/index.php/fm/article/view/13642`.

[97] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, and Y. Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. URL `https://arxiv.org/abs/2206.10789`.

[98] A. Q. Zhang, J. Amores, H. Shen, M. Czerwinski, M. L. Gray, and J. Suh. Aura: Amplifying understanding, resilience, and awareness for responsible ai content work. *Proc. ACM Hum.-Comput. Interact.*, 9(2), May 2025. doi: 10.1145/3710931. URL `https://doi.org/10.1145/3710931`.

[99] L. Zhang, X. Liao, Z. Yang, B. Gao, C. Wang, Q. Yang, and D. Li. Partiality and misconception: Investigating cultural representativeness in text-to-image models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642877. URL `https://doi.org/10.1145/3613904.3642877`.

[100] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng. Deep Long-Tailed Learning: A Survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10795–10816, 2023.

[101] D. Zhao, J. T. A. Andrews, O. Papakyriakopoulos, and A. Xiang. Position: measure dataset diversity, don't just claim it. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[102] N. Zhou, D. Bamman, and I. L. Bleaman. Culture is Not Trivia: Sociocultural Theory for Cultural NLP. *arXiv preprint arXiv:2502.12057*, 2025.

[103] A. Zur, E. Kreiss, K. D'Oosterlinck, C. Potts, and A. Geiger. Updating clip to prefer descriptions over captions, 2024. URL `https://arxiv.org/abs/2406.09458`.

A Figure Alt Text

Below, we provide alt text for Figure 1. For each cultural artifact (the braille notetaker and Mridangam), the figure displays one reference photograph and two AI generated images. We provide descriptions for each image below.

Braille notetaker:

- The reference photo for a braille notetaker shows a thin rectangular electronic device. The top surface of the device has a tactile braille display: one horizontal line of around forty braille cells. Above the display, the device has a braille keyboard and several other button controls. The device is sitting on a table, and a person's hands are resting on the tactile display.
- The first AI image shows two hands holding a device with a large black screen on its surface. The screen has small white dots that resemble braille, but are not tactile. The dots are arranged irregularly in about thirty columns and thirty rows.
- The second AI image shows a device shaped like a folding laptop computer, with an electronic screen display on top and a keyboard on the bottom. The screen of the device is displaying rows of small white dots that resemble braille, but are not tactile. The dots are also not arranged in valid cells. The keyboard of the device resembles a qwerty keyboard.

Mridangam:

- The reference image shows a Mridangam, a barrel-shaped percussion instrument, placed across the lap of a person seated cross-legged. It has two drumheads, one slightly larger than the other, and multiple thick strings laced along its body, connecting both ends. The playing surface on the visible side features a black circular area in the center, where the person's hand is positioned for playing
- The first AI image shows a person seated on a raised surface with a cylindrical, kick drum-like instrument placed horizontally on the floor in front of them. The instrument has a wooden finish playing surface facing forward, with a black circular area in the center. It also features strings running from one end to the other.
- The second AI image shows a close-up of a snare-like drum with a single drumhead on top and metallic rims around the edges. The drumhead is light-colored, with a small black pattern at the center.

B Study Protocol

In this section, we describe the study protocol (published elsewhere) that we followed to hold workshops with both communities.

B.1 Recruitment & Workshop Preparation

Our position paper draws from our experiences conducting participatory research engagements with two different communities: members of the blind and low vision ("BLV") community in the UK, and residents of Tamil Nadu or Kerala, two South Indian nation-states. These two communities have distinct histories and material cultures. However, past research has surfaced that frontier models fail to accurately depict both assistive technologies used by BLV community members [57] and South Asian cultural artifacts [72, 82].

Our study was approved by our institution's IRB. Community members were compensated an amount appropriate to their local context: £75 for those in the UK and Rs. 500 for those in India.

Designing an accessible protocol Our study protocol was largely similar across the two communities, but differed in two meaningful ways. First, we facilitated workshops with South Indian community members in *focus groups* with multiple community members to encourage dialogue and discursive evaluation [8, 73]. BLV community members participated individually, with a partner who they already knew. Second, we provided BLV participants with alt text descriptions for each image.

We follow past work [62] that invites blind community members and sighted partners to work together in pairs to evaluate AI-generated images. Past scholarship has demonstrated the benefits of "cross-ability collaborative work" [6, 92], e.g., where a blind user and sighted partner work together to complete a task. While past studies have shown that sighted strangers can misunderstand the access needs of their collaborators, recent studies have adopted cross-ability protocols between participants who already know each other well and have established trust and comfort working together [62]. Thus, we adapted our study activities (e.g., reacting to AI-generated images) to be formulated as a cross-ability dialogue between two participants. Each workshop with the BLV community consisted of 1 cross-ability participant pair.

When asking blind participants to evaluate AI-generated images, we began by providing an "alt text" description of what is shown. The alt text was written collaboratively with our co-author who was a community member. To encourage consistency in the amount of description given, our research team created an alt text "template" of important characteristics to describe for each image, following alt text best practices for photographs [62] (e.g., for each braille notetaker image, we always described its shape, material, surfaces, and any displays, buttons or keys). During the study, we also invited participants to further discuss what is shown in the images as a pair.

We began by presenting alt text prepared by the research team for two reasons. First, from a human-centric perspective, past studies that asked blind participants to evaluate AI-generated images found that participants found it difficult to ask questions about what was shown in the image and preferred to be provided with a basic description first [40]. Second, from an experimental design perspective, we wanted to standardize the information that participants were provided about each image so that we could better compare participants' reactions when presented with the same content.

Recruitment Like past participatory AI studies that center marginalized community members (e.g., [8, 57, 72]), we adopted a purposive sampling approach [65] to recruit members of both communities.

For the blind and low vision community, we recruited participants from two email lists: an internal list of blind and low vision community members who had consented to receive information about future studies at our institution, and an open list for blind and low vision technology users in the UK. We asked each community member to invite their sighted partner to the study, following Muehlbradt and Kane [62]. Relationships between blind community members and their partners included friends, partners, siblings, and children.

For residents of Kerala and Tamil Nadu, we sent out the calls for participation via X.com and also circulated it on WhatsApp. We received over 75 responses for both states combined. Based on community members' age bracket, familiarity with the culture, gender diversity, and their experiences/interaction with each artifact, we selected 5 community members to participate for each artifact.

Generating images We generated images using two T2I models that achieved state-of-the-art performance at the time we conducted our study: Stable Diffusion 3 Medium ("SD-3") and DALLE-3. SD-3 Medium was the most recent and most popular open source T2I model at the time we conducted our study, with over 3 million downloads on HuggingFace. DALLE-3 is a popular closed-source model and competitor to SD-3 that was widely in-use at the time of our study (e.g., is integrated into ChatGPT Premium). The images generated by these models reflect the state-of-the-art in the field, and are similar to those that participants may encounter "in the wild".

To generate images, we used two different prompt templates. The first template is a basic template ("*a photo of a {object}*") that is frequently used in past work [71, 74] to prompt for photorealistic images of objects in isolation. We also experimented with a few basic prompt engineering interventions (e.g., [17, 71]) to append descriptive text to each prompt. We explored several strategies:

1. Append a suffix with a description of the community, e.g., "*a photo of a braille notetaker used by someone who is blind or low vision*"
2. Append a description generated by GPT-4, following Pratt et al. [71], e.g., "*a photo of a braille notetaker. [LLM-generated description]*"
3. Append a handwritten description of important visual components of the object, written by community members on our team. e.g., "*a photo of a braille notetaker. A braille notetaker used by people who are blind is an electronic device shaped like a rectangle. The top of the device has 8 round buttons and a space bar. The bottom of the device has a tactile braille display with braille cells. Each braille cell on the display has 8 holes where tactile braille pegs can come out. The sides of the device have other buttons and ports to charge the device.*"

Following past T2I user studies [57], we pre-generated AI images so that synchronous study time could be spent eliciting participants' feedback. Because participants only had time to review a small number of images (shown one-at-a-time), the research team curated sets of images that reflected meaningfully different depictions of each object. Images were chosen by the last author and a community member on research team.

B.2 Workshop Activities

All workshops involved an activity where participants were invited to annotate and respond to AI generated images.

We began the study by introducing the activity, and invited participants to reflect on their familiarity with and knowledge about the artifact being evaluated.

Q1. Is this image a correct depiction of a [object]? (OUTPUT: 1-3)

1. The image is wrong – it is totally unlike the object
2. The image is partially correct – some aspects are correct
3. The image is correct

Q2. What are the 3 most important things that would have to change, for this image to be a correct depiction of a object? (OUTPUT: Free text, three items)

ALT: What is good and what is bad about this image?

Q3. Can you rank whether you prefer this image, in comparison to the other image(s) from before? (OUTPUT: Ranking order)

Q4. Why did you put it there? (OUTPUT: Free text, 2 sentences)

Q5: Say that you were using AI to add images to a presentation to share with your (class, workplace, friends) about your summer holidays. Which of these five images would you be worried about your peers seeing?

ALT: Did you find any of the images shown to be offensive or upsetting?

C Scaling Community Annotations: Extended

Standard error approximation for Spearman's rank correlation We use Bonett's approximation for the standard error [11, 26]: $SE_\rho \approx (1-\rho^2)/\sqrt(n-3)$. We can then solve for the number of samples n by rearranging the formula as: $n \approx ((1-\rho^2)/(m/z))^2 + 3$, where ρ is Spearman's rank correlation coefficient (e.g. 0.7), m is the desired margin of error (e.g. ± 0.05), and z is the z-score corresponding to the desired confidence level (e.g., 1.96 for 95$\%$ confidence).

D Shades of Bad: Extended

Object	Total Annotations	1 (Wrong)	2 (Partially Correct)	3 (Correct)
Braille notetaker	15	8 (**53%**)	5 (**33%**)	2 (**14%**)
Mridangam	80	67 (**84%**)	13 (**16%**)	0 (**0%**)

Table 1: Distribution of rating annotations assigned to generated images. Three annotators were shown 5 images of a braille notetaker. Five annotators were shown 16 images of a Mridangam.

Object	img1	img2	img3	img4	img5
Braille notetaker	1.67	1.67	1.67	2	1

Table 2: Ratings averaged across 3 annotators for the 5 images of a braille notetaker.

Object	Ratings
Mridangam	[1.2, 1.8, 1, 1, 1.2, 1.4, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1]

Table 3: Ratings averaged across 5 annotators for 16 images of a Mridangam.

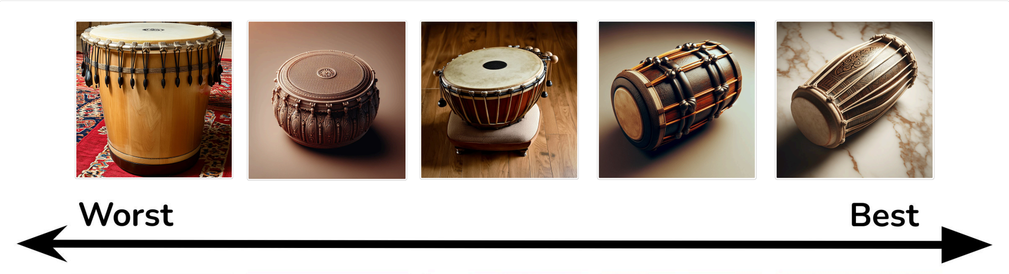

Figure 4: **Final ranked images fo Mridangam via group consensus.** Although participants initially ranked the images individually from best to worst, reflecting variation in their preferences, the final set of rankings was determined through group deliberation and consensus.

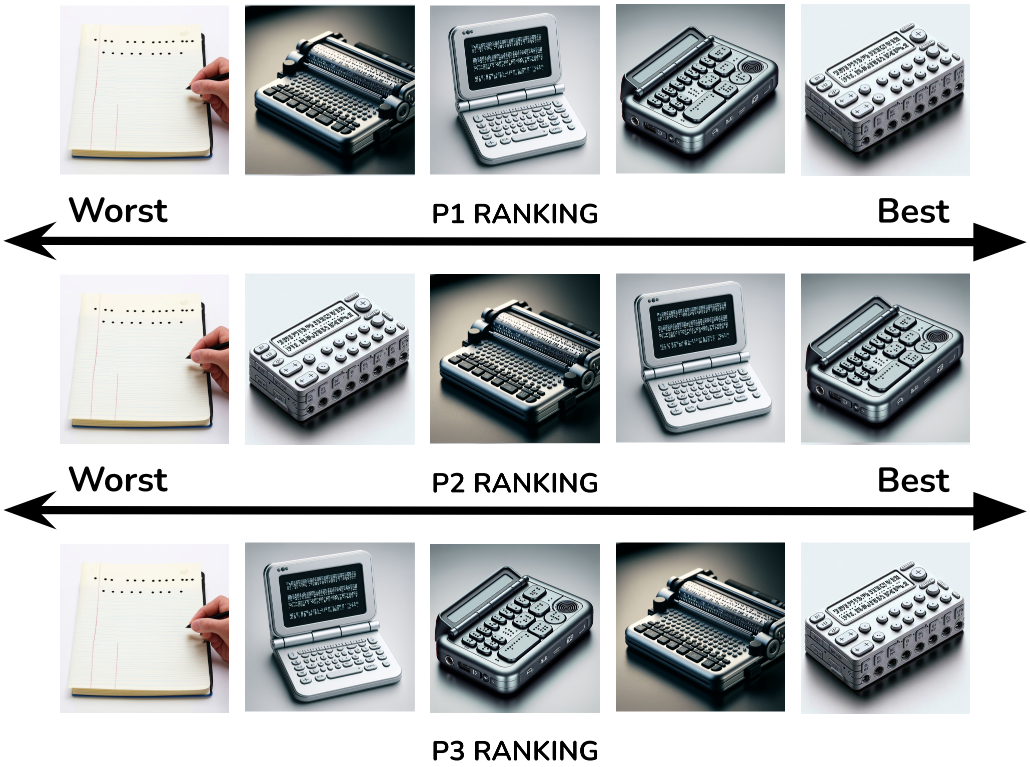

Figure 5: **Variance in participants' ranking preferences for images of a braille notetaker.** While all three participants consistently agreed that depictions of notetaking use paper were their least favorite, we observed high variance in how participants ordered other inaccurate depictions of a braille notetaker (e.g., images that depict a braille notetaker similarly to a handheld calculator, manual typewriter, or laptop computer). This variance reflects the difficulty of making arbitrary judgments between "shades of bad".