# Engaging Communities Meaningfully in Defining Disability Representation for AI Image Generation


Anja Thieme
Microsoft Research
Cambridge, UK
anthie@microsoft.com

Rita Faia Marques
Microsoft Research
Cambridge, UK
t-ritaf@microsoft.com

Martin Grayson
Microsoft Research
Cambridge, UK
Martin.Grayson@microsoft.com

Sidhika Balachandar
Microsoft Research
Cambridge, UK
UC Berkeley
Berkeley, USA
sidhikab@berkeley.edu

Cameron Tyler Cassidy
Microsoft Research
Cambridge, UK
University of California
Irvine, USA
cameron.cassidy@uci.edu

Madiha Zahrah Choksi
Microsoft Research
Boston, USA
Cornell University
New York, USA
mc2376@cornell.edu

Camilla Longden
Microsoft Research
Cambridge, UK
Camilla.Longden@microsoft.com

Reeda Shimaz Huda
Microsoft Research
Boston, USA
Georgia Institute of Technology
Georgia, USA
rhuda8@gatech.edu

Nicholas Ileve Kalovwe
Kilimanjaro Blind Trust Africa
Nairobi, Kenya
nicholas@kilimanjaro-blindtrust.org

Christina Mallon
Microsoft Corporation
Miami, USA
cmallon@microsoft.com

Courtney Mansperger
LPA
Chicago, USA
courtney.mansperger@lpaonline.org

Daniela Massiceti
Microsoft Research
Sydney, Australia
daniela.massiceti@gmail.com

Bhaskar Mitra
Microsoft Research
Montreal, Canada
bhaskar.mitra@gmail.com

Ruth Mueni Nzioka
Short Stature Society of Kenya
Nairobi, Kenya
ruthmueninzioka@gmail.com

Ioana Tanase
Microsoft Corporation
Toulouse, France
ioanat@microsoft.com

Yuzhe You
Microsoft Research
Cambridge, UK
University of Waterloo
Waterloo, Canada
y28you@uwaterloo.ca

Cecily Morrison
Microsoft Research
Boston, USA
cecilym@microsoft.com



## Abstract

Media representations of people with disabilities profoundly influence societal perceptions, yet have historically been absent, stereotyped, or inaccurate. As AI-generated visual media becomes increasingly prevalent, there is a critical opportunity to address these misrepresentations. Responding to the lack of collectively negotiated representation standards, this paper presents our human-centric approach to engaging disability communities meaningfully in AI data practices. Over three months, we worked closely with three disability organizations across the Global North and South to develop the Community Library Creator that introduces design scaffolds to support communities in defining 'good' representation and curating community-centric AI datasets; laying the foundations for community-specific evaluation metrics and future model adaptations. We contribute qualitative insights into the complexities of community-led data curation; discuss the value and practical challenges of intersecting human insights with AI requirements;




*CHI '26, Barcelona, Spain*



ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790768

and reflect on human-centered AI approaches that empower communities to share their perspectives and actively shape AI data practices.

## CCS Concepts

• **Human-centered computing → Participatory design**; **Empirical studies in HCI**; • **Computing methodologies → Machine learning**.

## Keywords

Disability representation, human centered AI, human participation in AI pipelines, community participation, empowerment, generative AI, GenAI, text-to-image models, equitable AI, inclusive design

**ACM Reference Format:**

Anja Thieme, Rita Faia Marques, Martin Grayson, Sidhika Balachandar, Cameron Tyler Cassidy, Madiha Zahrah Choksi, Camilla Longden, Reeda Shimaz Huda, Nicholas Ileve Kalovwe, Christina Mallon, Courtney Mansperger, Daniela Massiceti, Bhaskar Mitra, Ruth Mueni Nzioka, Ioana Tanase, Yuzhe You, and Cecily Morrison. 2026. Engaging Communities Meaningfully in Defining Disability Representation for AI Image Generation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3772318.3790768

## 1 Introduction

How people with disabilities (PwD) are presented in media plays a pivotal role in shaping public perceptions and social norms [34, 105], influencing livelihood through access to education and employment [47]. Yet, media portrayals of PwD have historically been absent, stereotyped, harmful or inaccurate rather than reflecting authentic, lived experiences [34, 49, 88, 114, 117]. Although visibility of PwD is slowly increasing, one billion PwD globally remain largely invisible or inaccurately depicted in the media [5]. As AI-generated visual media becomes rapidly more accessible to people [81], there is opportunity to address this historical misrepresentation and promote a more authentic, positive representation of disability.

AI media generation, specifically image generation models that translate free-form text prompts into visuals [87, 94, 95], have been trained on datasets that reflect this historical erasure and bias. As a consequence, they currently depict PwD in unrealistic or dehumanizing ways [48, 77], and often omit or distort representations of assistive technologies [77]. Generative AI outputs tend to reproduce normative views and mirror societal biases for marginalized groups [15, 21, 36, 41, 42, 45, 53, 66, 72, 91], including PwD [10, 12, 39, 48, 54, 62, 77, 113]. Unlike other forms of marginalization (e.g., gender, race), disability representation has yet to be publicly debated or defined (cf. evolving work by [103, 117]). This calls for disability communities to lead the conversation in shaping a shift to positive representation as AI-generated visual media continues to scale.

To meaningfully shift representation in AI systems, we must reconfigure how disability communities are involved in AI data practices. Current data collection processes prioritize scale over inclusion, relying on web-scraping (e.g., [40, 97]) and synthetic pipelines that reflect developer biases and the uneven availability of online content [13, 14, 57, 96]. These approaches produce noisy and skewed datasets; they exclude communities from defining how they wish to be represented. Community involvement, if present, has often been limited to data collection, annotation or model output evaluation. Conceptualized representation and its translation in data are under-supported. This gap reinforces extractive dynamics and misses the opportunity to embed community expertise into the foundations of AI development. Despite calls for more inclusive data-centric AI (e.g., [11, 31, 53, 96, 107]), approaches that enable communities to define and curate datasets remain underexplored.

In this work, we partner with disability communities to meaningfully engage them in AI data practices through their defining and evaluating what 'good' representation means to them, and its embedding within a curated dataset – the *Community Library*. Over a period of three months, we work closely with three non-profit organizations representing *people with dwarfism* and *people with vision impairments* across the Global North and South to develop a first technical prototype—the *Community Library Creator*. Through this, we explore how technology scaffolds can help lower barriers to participation and embed community voice and values into AI data practices. Our approach lays the groundwork for disability community-specific evaluation metrics and future model adaptations towards the larger vision of improving representation in AI image generation.

Against this backdrop, our work makes three main contributions:

- We present our technology-supported, community-led approach that introduces various design scaffolds – an image-first workflow, a structured Community Library, community-centric prompt generation, and AI evaluation – to aid communities in specifying their desired representation and embed these insights into AI-interpretable data formats. We describe rich qualitative insights that highlight the individualized, dynamic, and negotiated nature of each communities' representation definition, and the complexities of community-led data curation such as balancing lived realities with future aspirations, and navigating intra-community diversity.
- We discuss the value of our approach and surface practical challenges for intersecting human insights with technical AI requirements (e.g., concept completeness, balance, or data scale). By surfacing these tensions, we outline future directions in this emerging research area for facilitating community-centric AI data practices.
- We reflect on human-centred approaches to empowering communities in AI data work, proposing (i) a shift from identifying harms to cultivating meaningfully representation; and discussing (ii) the use of advocacy organizations as proxies for bounded communities, and as infrastructures for community outreach and deliberation.

## 2 Related Work

This section outlines the: (1) importance of appropriate and inclusive disability representation in media, and existing efforts in its definition; (2) limitations of current image generation models, and the value of community-led evaluation to drive improvements; and (3) how existing data-centric AI practices underscore the need to involve (disability) communities in shaping those practices.

## 2.1 Disability Representation in (Digital) Media

Visual media plays a significant role in shaping public opinion and social norms, including societal perceptions of disability [10, 34, 105, 117]. It also impacts how PwD form their identity: positive portrayals can affirm identity, while negative ones may lead to denial [120]. Yet, PwD are frequently absent or negatively stereotyped in media [34, 49, 103, 120] – portraying them as objects of pity, charity, or humor [49, 88], or as super-heroes overcoming insurmountable barriers [34, 100, 120]. These framings typically reflect how disability is perceived by others rather than capturing the lived experiences of PwD, resulting in inauthentic or incomplete depictions [49]. Yet, there remains limited understanding of what it means to represent PwD and their communities well in (digital) media to drive meaningful change.

Several works have explored how PwD wish to represent themselves in digital depictions, mainly as virtual avatars or agents (e.g., [9, 33, 43, 76, 118, 119]). Existing research found that positive representation requires empowering users to control how and when they disclose their disability – whether by showing physical differences, assistive technologies (ATs), or through behaviors and symbols like movement patterns, sign language, or cultural icons [76, 119]. Furthermore, disclosure varies by person and context: some people, and depending on the social context, embrace disability as a core identity, while others generally or temporarily minimize it to *"present a capable self"* [118]. As such, the literature suggests flexibility, providing a range of potential depictions [76, 118, 119].

In this paper, we shift focus from individual self-portrayals and personal preference to 'collective representation' by exploring how disability communities negotiate their preferred representation to audiences external to their community.

## 2.2 Community-Defined Concepts and Data for AI Evaluation & Image Generation

Previous research has highlighted challenges for image generators to depict PwD in respectful and realistic ways [48, 77], echoing broader biases seen in representations of other marginalized communities [12, 21, 81, 91]. Mack et al. [77] conducted focus groups examining how PwD were represented in image generation models. Their findings revealed a recurring reliance on reductive archetypes – portraying PwD as sad, lonely, inactive, or incapable. Additionally, assistive technologies (ATs) were frequently foregrounded over the individuals themselves, leading participants to describe such imagery as 'dehumanizing' [77]. These findings highlight how societal stereotypes manifest in AI systems and underscore the need to address representational harms for disability communities.

The limitations of current image generation models in representing PwD stem from the data pipelines and practices typically used to train them. Web-scraped data for AI model pre-training is skewed towards an underrepresentation of disability data [69, 79]. Even when PwD data is available, data from marginalized groups is often automatically filtered out by techniques for removing 'low' quality data [57, 104], which rely on biased models such as CLIP [92]. Addressing these knowledge gaps through large-scale model *pre*-training is challenging due to the scale of the data required, computational resource needs, and the proprietary nature of many advanced models. Nevertheless, smaller datasets can be vital in enabling key *post*-training techniques such as fine-tuning, in-context learning, and prompt engineering (cf. [86, 122]).

Data alone is not enough to improve AI models. It is important to evaluate whether AI models produce respectful and accurate representations of disability communities [65]. However, common image generation evaluation metrics — such as CLIPScore [55], FID [56], and uses of multimodal large language models (MLLMs) 'as-judge' metrics [58, 71] often penalize authentic depictions of ATs, while rewarding incorrect or even offensive renderings (e.g., Braille displays shown as paper) [65]. They also tend to over-index on superficial realism and lack the domain-specific knowledge needed to assess nuanced aspects of disability representation. Taking a different perspective, recent social science work [89] has proposed richer evaluation categories that can give more nuance to assessments of cultural images, including: incorrectness, missingness, specificity, coherence, and connotation. This growing body of work underscores that evaluation approaches are in their infancy, and that solutions must be both technically practical and socially situated.

A particularly promising evaluation approach is fine-tuning MLLM 'judges' on context-aware or community-sensitive data [75, 111]. A noted current challenge is the required foundational step of: clearly defining what constitutes 'good' representation for a community that can be used in the fine-tuning process. Recent calls – in generative AI evaluation frameworks [115] – to separate the systematization of the concept to be measured, and the operationalization of the measure technically,further emphasizes the need to articulate what is evaluated before determining how to evaluate it.

Our work contributes to this emerging area by demonstrating how technology scaffolds and direct community-engagement can support this important definition step and enable data collection to operationalize meaningful metrics and evaluator models for assessing AI-generated images of disability communities.

## 2.3 Bringing (Disability) Communities into AI Data Practices

The domain of Human-centered AI HCAI calls to actively involve people in AI design to ensure systems better reflect the values, preferences and needs of users, and other impacted stakeholders (e.g., [4, 8, 16, 99, 101, 104, 110]). However, in a recent review paper, Delgado et al. [31] found that most AI projects bring stakeholders merely in as consultants, in one-off preference elicitation or UI decisions (cf. also [27, 108]). Similarly, in research involving PwD, participants mostly collaborate on the design or evaluation of access-supporting generative AI systems [2, 3, 22, 60, 61, 73, 98, 109]. There is also a growing trend in proxy-based participation, whereby individuals familiar with a stakeholder community, including UX/HCI practitioners, often 'stand-in' for others [27, 31, 106] rather than enabling communities to speak for themselves. Beyond AI system design and evaluation, there is a need to bring people *meaningfully* into data practices of defining relevant concepts and data for AI, and in directing the evaluation of AI outputs.

*2.3.1 Community-centric AI Data Curation.* To address representational gaps and biases in AI models, a number of community-centric datasets are emerging aimed especially at improving cultural and geographical representation (e.g., DOSA [99], World Wide Dishes [53,

78], or GeoDE [93]); and language diversity (e.g., Masakhane [84], Aya [102]). In disability contexts, key datasets include ORBIT [80] – videos and images of objects recorded by people who are blind or low-vision; StammerTalk [74] and ASL Citizen [32] for speech and language impairments; and the grassroots Disabled And Here stock image collection celebrating disabled Black, Indigenous, People of Colour (BIPOC) [6]. These existing works draw attention to the larger *data work* that characterizes the creation of high-quality, community-centric datasets [53], pertaining to: (1) the labor involved by 'participatory mediators' in dataset construction, who are crucial for building trust and rapport with community members [53]; (2) the importance of information scaffolds and other educational resources to making participation accessible [31, 53, 107]; and (3) the need to contextualize community values to support meaningful data collection [11, 53, 90]. This requires recognizing that data is not only relational, but also shaped by its creation context, whereby datasets reflect the worldviews and beliefs of their creators who decide what deserves capturing and how information is classified [17, 53]. Furthermore, Qadri et al. [90] advocate for new data curation and annotation methods that center interpretation and deliberation of data's social meanings, for example, through expert workshops for socially contested concepts (e.g., disability representation) – *aiming to construct and debate data needs and labels more collectively*. This underscores how dataset curation involves more than 'technical' work. It raises questions about who participates in these debates, whose values and norms should shape AI, and how community-centric deliberation in data definition and curation can best be supported [11].

*2.3.2 Human Involvement in AI Data Annotation and Evaluation.* Where humans annotate data for AI image evaluation, their role commonly involves: identifying objects or features in images [52]; image comparisons to indicate similarities or preferences across two or more samples (e.g., [52, 85]); or gamified approaches (e.g., describing artifacts characteristics for another annotator to guess [99]). The most dominant approach to gathering human annotations is to 'crowd-source' [7, 53, 65] either anonymous people or community members as 'data workers', who are given simple, tokenistic tasks "that even inexperienced annotators can finish in an instance" [31, 85], with little or no context to the data. In expressing 'their' perspective through data annotations, it important to recognize how the identity and views of human annotators become embedded in the data, influencing subsequent system behavior [51, 52, 90]. For example, Hall et al. [52] found that annotators, who were not from the geographical region depicted in AI images tended to favor exaggerated, stereotypical images and overlooked other realistic, representative portrayals. This has led to calls for more inclusive, community-centric AI evaluation approaches [11, 68] that leverage qualitative methods like focus groups [89] and online workshops [68] to capture nuanced, socially situated and dynamic aspects of concepts like cultural or disability representation; as well as solicit community-proposed prompts to ground AI evaluations in community relevance (e.g., [10, 77, 89]).

*2.3.3 AI Data Stewardship.* Data stewardship is defined as the *"responsible use, collection and management of data in a participatory and rights-preserving way, informed by values and engaging with questions of fairness."* (p.4) [63]. Data stewards – individuals or organizations – govern data on behalf of beneficiaries [1, 64], involving community members and safeguarding community-specific materials by defining their use terms [25, 63] (e.g., via licensing). Examples include data trusts or cooperatives, which use legal frameworks to formalize collective data management such as data commons [30, 112] or community archives. *AI data stewardship* applies this approach to dataset collection for training AI models, especially when these have knowledge gaps about marginalized communities (cf. [20, 59]). In this paper, AI data stewardship is about meaningfully engaging disability communities in defining their own representation for AI, and ensuring that data practices are participatory and aligned with community values. Details on our data governance specific work are reported elsewhere [25].

## 3 Method

This section outlines (1) our approach to recruiting disability organizations; (2) how we engaged disability organizations in AI data practices through the design of the Community Library Creator prototype in conjunction with broader community engagements; and (3) our methods for data capture and analysis.

### 3.1 Disability Communities & Project Leads

We work with disability advocacy organizations as proxies for bounded disability communities. Such organizations have defined memberships and existing structures of communication and negotiation to support deliberations for defining visual identity. This approach also sets appropriate boundaries for determining the level of diversity needed in imagery to represent a community and, most importantly, avoids assuming a shared identity based solely on disability. For instance, blind people in Kenya may have different views (e.g., what do I want the world to see) and visual identity (e.g., different ATs) compared to blind people in the United States. We acknowledge that power dynamics between advocacy organizations and their members may vary.

We recruited three disability communities via internal connections within Microsoft, with outreach coordinated through an international NGO. We focused on disability organizations that specifically represent *people with dwarfism* and *people with vision impairments* as two disability groups for whom internal evaluations[1] showed image generation model outputs to be particularly poor. While both groups reflect 'visible disabilities' (vs. neurodiversity), they pose distinct challenges for AI: visual characteristics of dwarfism are inseparable from the person, whereas the identification of someone who is blind or has low vision often requires external identifiers (e.g., a guide cane). Further, we included communities that span geographically and culturally diverse locations to ensure our technology-supported, community-centric process can adapt across requirements of the Global North and South.

Each organization signed a contract that included financial support for giving time and focus to the research, asking to: partake in research activities; curate a Community Library of 400 images

[1] A manual, human review by two disability experts of images generated for 662 prompts (612 benign, 50 adversarial), prompted across ten main types of disabilities (e.g. mental health, blindness, low vision, mobility, neurodiversity, learning, speech, deafness, hard of hearing, pan disability) using an off-the-shelf image generation service found that defect rates were higher for Dwarfism and Blind and Low Vision communities.

| | |
|---|---|
| **Organization 1**<br>**Location:** Global South (Kenya) | |
| **Project Lead:** P1 | **Disability community:** People with Vision Impairment. |
| **Duration at organization:** 3 years | **Mission:** Ensure that children and young adults with visual impairments achieve academic success and social, economic inclusion through quality education & state-of-the-art ATs. |
| **Age:** 30 | **Members & Reach:** Over 3,000 learners in 250 schools and colleges across Kenya. |
| **Gender:** Male | **Role within organization:** Supports programs that focus on co-designing, testing, and iterating Assistive Technology (AT) solutions, including working with AT innovators to pilot and refine digital technology applications that are tailored to the specific needs of persons with disabilities. |
| | **Relationship with org members:** Has well-established, trusted and reciprocal relationship with a diverse, especially tech-savvy professional community; sharing research findings and discussing challenges. |
| | **Individual project motivation:** Driven by a passion for supporting PwD and making a lasting, positive impact on their lives; wants to learn more about AI. |
| **Organization 2**<br>**Location:** Global South (Kenya) | |
| **Project Lead:** P2 | **Disability community:** People of Short Stature. |
| **Duration at organization:** 12 years | **Mission:** Aims to end discrimination towards people of short stature and promote their rights in all aspects of community life. |
| **Age:** 42 | **Members & Reach:** Over 600 members. |
| **Gender:** Female | **Role within organization:** One of the founders of the organization. Started the organization by participation in a beauty pageant and gaining media attention in 2013. |
| | **Relationship with org members:** Members are employed in formal professions and engage primarily through social media and in-person gatherings, since many are not tech-savvy. This government-funded organization offers capacity-building programs and mental health support, delivered through home visits and group events (e.g., sports events). |
| | **Individual project motivation:** Aims to empower members to shape their own representation, recognizing the need to include their community in technology development and improve inclusion more broadly (e.g., accessibility and product design). |
| **Organization 3**<br>**Location:** Global North (USA) | |
| **Project Lead:** P3 | **Disability community:** People with Dwarfism. |
| **Duration at organization:** 1.5 years | **Mission:** Improve quality of life for people with dwarfism throughout their lives while celebrating with great pride little peoples' contribution to social diversity. |
| **Age:** 30 | **Members & Reach:** Over 7,500 members. |
| **Gender:** Female | **Role within organization:** Administrative manager; mostly outward facing role, supporting anybody who has questions about the organization (by call, e-mail). |
| | **Relationship with org members:** The organization operates on a paid membership model managed by volunteers. Its members, primarily families, are highly engaged, especially through in-person events and conferences; serving as the leading resource hub for the dwarfism community in the USA. |
| | **Individual project motivation:** Opportunity to improve how dwarfism is portrayed in the media and celebrate the diversity in the community while personally learning about AI and engaging more deeply with the organization. |

**Table 1: Overview of the three disability organizations including their mission, members and reach as well as the demographics, role and project motivation of the respective project leads (P1-P3).**

with corresponding annotations; and open-source the final dataset for greater reach and impact. Table 1 provides an overview of each organization, and the project lead who volunteered to partake as the primary participant in this research – advocating for, and actively mobilizing the voice of their community members in the process. Project leads are given a unique identification number to protect their anonymity (P1-P3). The research study was carefully reviewed for compliance and IRB approved (Reference ID: 11019). Informed consent was sought in writing prior to the study.

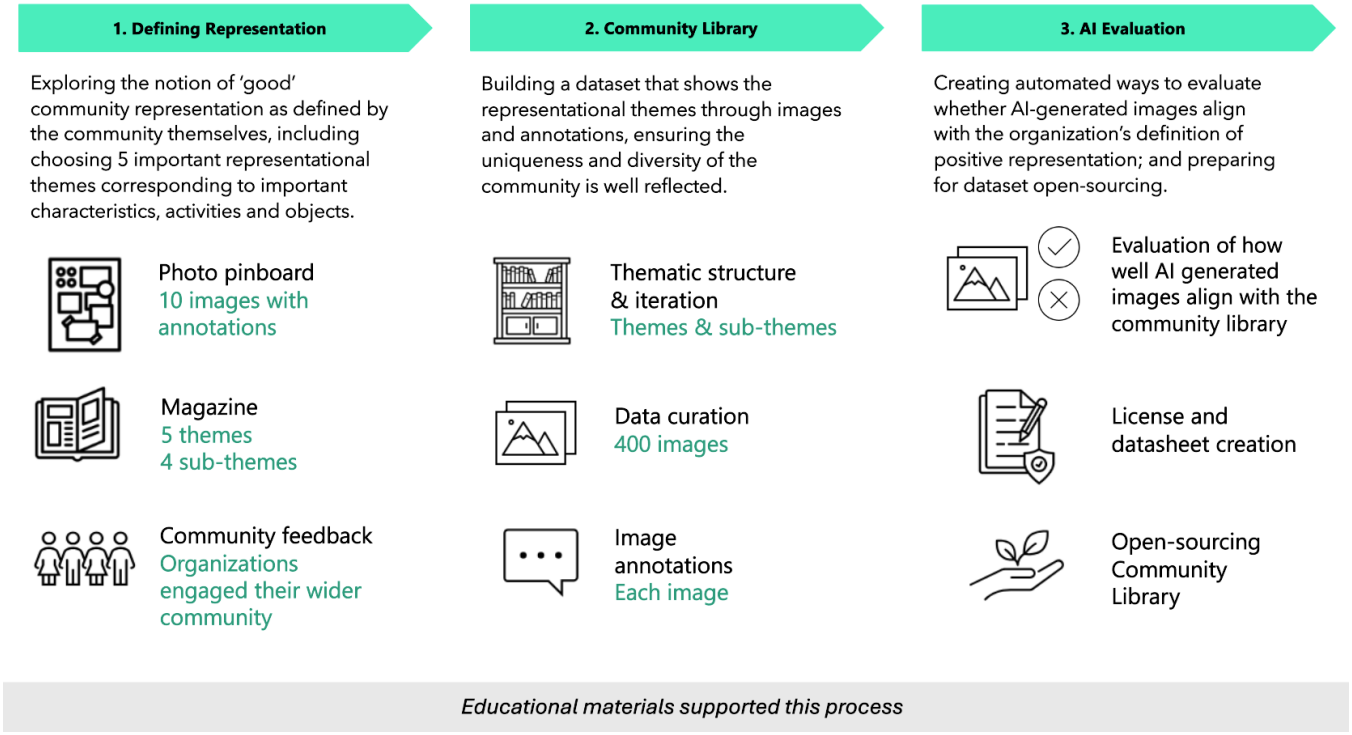


**Figure 1: Overview of the three consecutive phases of our technology-supported, community-centric engagement process.**

## 3.2 Our Technology-Supported, Community-Centric Engagement Process

Aiming to lower barriers for non-AI experts to participate and enable communities to embed their own values, choices and language in AI data practices, we developed a first prototype of the Community Library Creator. This prototype provides various reflection scaffolds and data structures to aid disability organizations through a three-phase process (see Figure 1) of: (1) *defining 'good' representation*; (2) creating a *Community Library* of 400 images with relevant annotations for AI; and (3) enabling the development of community-centric *AI evaluation* metrics. Initially, only the first activity of Phase 1 was prototyped. The remaining technical components evolved progressively using agile methods. Each week, our team of HCI, design, and AI researchers collaborated to translate research insights that evolved through our engagements with the three disability organizations into new features. This involved running machine learning (ML) experiments in parallel to clarify design requirements, and vice versa, adjusting ML approaches based on realistic user inputs. Our prototype is built using ReactJS for the front-end interface and a FastAPI server hosted on Azure. Data is stored in a SQLite database with image data stored in Azure Blob Storage. Our approach was guided by three main design principles, aiming to:

- create an empowering experience by prioritizing for disability communities to define ***positive*** representation rather than discussing any harmful, stereotypical depictions, or exposing them to negative AI outputs.
- enable a plurality of disability communities to have agency in bringing in their own values, definitions and representational choices through their specification and curation of a ***community-centric*** dataset.
- balance a community-centric experience with the technical constraints of making human insights and data ***AI interpretable*** without being limited to any specific image generation model or use context.

Over three months, project leads from each disability organization committed three days per week to: (i) group-based, educational webinars (see Supplementary Materials for an example) as well as reflective, results sharing meetings to foster dialogue and the exchange of ideas and leanings among project leads; (ii) individual research activities including interviews, platform use, and support sessions; and (iii) self-organized community outreach, explaining the project to members to solicit their input and engaging them in capturing or collecting community images. Next, we detail the main activities of each engagement phase. See Figure 2 for an overview.

*3.2.1 Phase 1: Defining Positive Representation.* Project leads were supported in defining what positive visual representation means for their community through three sequential activities:

**(1) Photo Pinboard** – Asking project leads to articulate a desired representation of their community can be challenging. In initial explorations amongst research colleagues, we found that responses about community aspirations or values can be very abstract and cognitively demanding; lack specificity of what is meant by concepts like independence or diversity; and also clarity of how

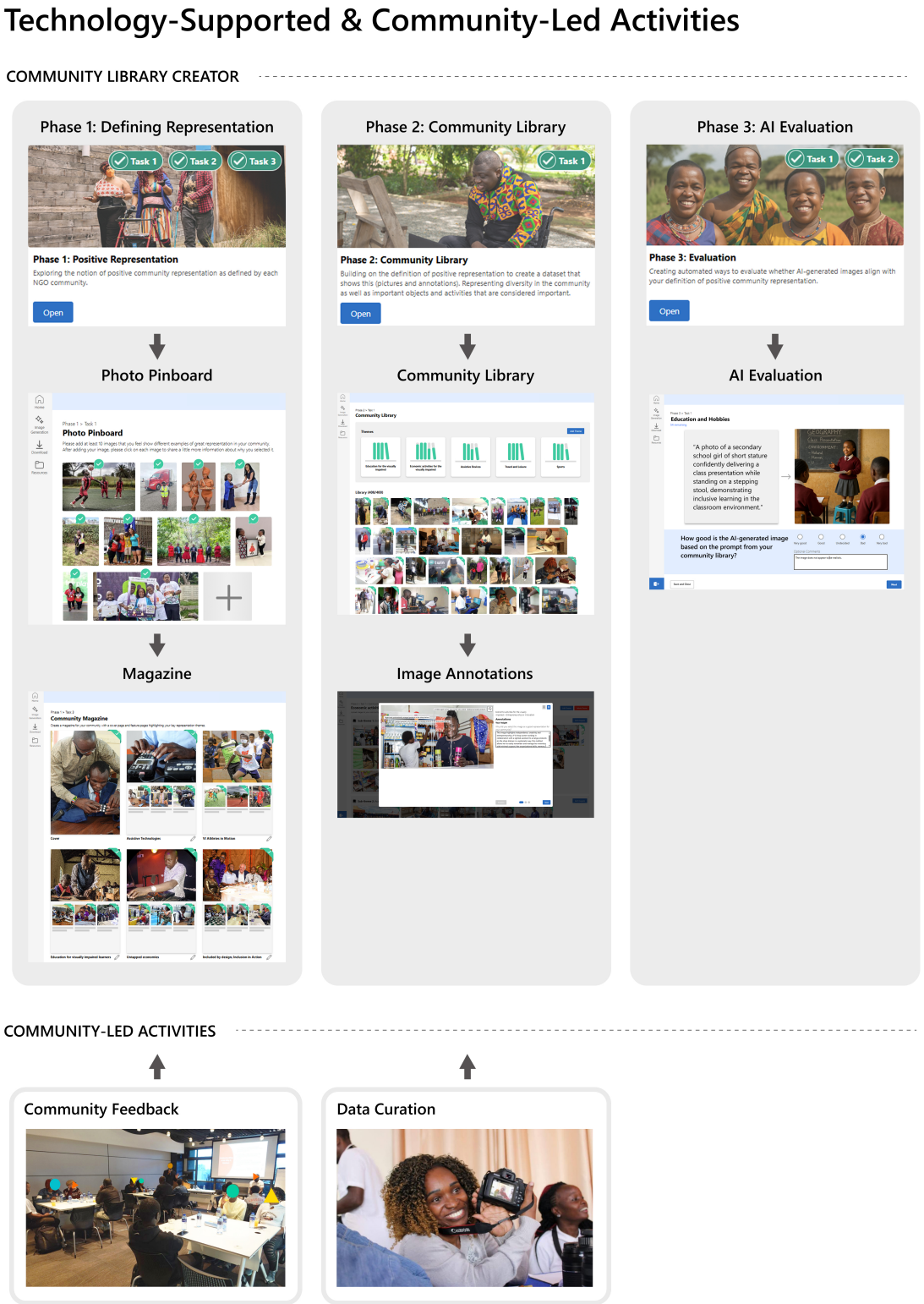


**Figure 2: Overview and flow of the technology-supported and community-led activities spanning the three consecutive project phases. The main interface includes images of the Disabled And Here collection [6].**

these appear visually. To address this, we chose an *image-based* approach, asking project leads to start their definition process by selecting 10 pictures that represented their community well – using the metaphor of a Photo Pinboard (Appendix A.1, Figures 6 and 7). Using each image as visual reference, they are then invited to reflect on three questions to surface initial representation themes, activities or aspirations: Why did you select this image? What makes this a 'good' representation of your community? What do you want others to understand about your community through this image?

**(2) Magazine** – Leveraging the metaphor of a Magazine (Appendix A.1, Figure 8), the second activity was designed to help project leads narrow down and prioritize the most important representation themes that evolved through their Pinboard, and reflect on how their community would like to be perceived in the world. This step was motivated by AI evaluation processes requiring well-defined concepts to evaluate for success. Therefore, to extract 'key' themes and invite reflections about important representational 'variances' within each theme, the activity asks project leads to select a cover image and title, followed by five feature pages – each structured into one main and three supporting visuals. If required, they have the option to add additional feature pages.

**(3) Community Outreach** – Recognizing the influential role of project leads in defining community representation, it is essential to also involve community members directly into the definition process. As advocates embedded within their community, project leads self-organized their outreach using familiar methods for engaging with their members. We supported their processes through planning and reflective questions to understand: (i) how they mobilize their communities and ensure diverse representation; and (ii) how they gather, deliberate, and ultimately integrate stakeholder input into their evolving community representation definition.

*3.2.2 Phase 2: Curating the Community Library & Data Annotations for AI.* In Phase 2, project leads translate the key representation themes from Phase 1 into a structured image dataset – the *Community Library* — with *image annotations*.

**(1) Community Library** – The Community Library provides a 'thematic structure' to aid project leads to go from initially defined higher-level representation aspirations of their Magazine to more specific instantiations (Appendix A.1, Figures 9 and 10). This 'thematic structure' was determined by the needs of downstream model training and evaluation, specifying that: (i) themes comprehensively include all contexts that matter to the community; (ii) each theme contains a similar number of sub-themes that span the full range of relevant contexts within that theme; and that (iii) each sub-theme contains a roughly equal number of images whilst, as a collection, it covers how a community wishes to be represented in that sub-theme. This 'balanced' structure is necessary to prevent data skew or bias, where overrepresented themes or sub-themes could disproportionately influence the behavior of models trained on this data. It also enables controlled comparisons across themes and sub-themes, supporting disaggregated and more interpretable evaluation. To curate a balanced 400-image dataset – assuming five themes and four sub-themes – allocates ~20 images per sub-theme. We chose 400 images to minimize burden of community-led curation while prioritizing quality over quantity.

**(2) Image Annotation** – Image annotations, including prompts and spatial cues [92, 121], are essential for training and evaluating image generation models as they provide localized and semantically rich supervision that help models learn meaningful visual representations. After iterating on annotation design, we opted to ask project leads to describe each image by answering the question: 'why' they had chosen to select an image as a good representation; and to highlight — using labeled bounding boxes — up to five *objects* and *people/animals* (e.g., guide dog) that are of particular relevance to their community. In addition, to produce representative AI images for evaluation purposes, we opted to auto-generate community-relevant 'image-level prompts' from project leads' response to the 'why' question and the corresponding image, using GPT-4o [83]. See Figure 3 for an overview of the three annotation steps, and Appendix A.1 Figures 11-13 for details.

*3.2.3 Phase 3: Evaluating AI.* Finally, Phase 3 grounds each community's representation definition in the outputs of the latest image generation models to better align the data with downstream model alignment and evaluation needs. To achieve this, we ask project leads to assess how well AI-generated images reflect their community representation goals through a rating task. This task invites a holistic judgment of each image – an approach chosen for its simplicity, scalability, and ability to capture community preferences in a structured yet open-ended format. Ratings also serve as a core input to model alignment pipelines, underpinning approaches like reinforcement learning from human feedback (RLHF) [26], and forming the basis of large-scale evaluation platforms such as LM-Arena [23].

For this evaluation task, project leads are shown images generated from the image-level prompts of their Community Library and asked to rate how 'good' the image is given the prompt on a 1–5 scale (1 = very bad, 5 = very good), with comments optional to reduce burden (Figure 5). To generate the images, we used: GPT-Image-1 [82], Imagen4-Ultra [50], Stable Diffusion 3.5 Large Turbo [35]. For each community, we randomly selected five prompts from each of ~20 sub-themes in their library, resulting in 100 prompts as inputs to the three image generation models. We filtered out any clearly offensive or harmful content (e.g., infantilization for dwarfism) to focus community feedback on nuanced aspects of representation, yielding ~300 images per community. Resulting ratings are intended to serve as community 'preference signals' for influencing potential downstream tasks such as: training 'evaluator models' that automatically assess whether new AI-generated images reflect a community's representation preferences; or steering image generation models towards community-preferred depictions.

## 3.3 Research Data: Capture & Analysis

All research meetings were recorded and transcribed via Microsoft Teams. Key segments were checked for correctness before transcripts were anonymized and recordings deleted. The data corpus also includes: (i) any content that project leads uploaded to the prototype – time-stamped and linked to a user ID; (ii) post-study questionnaires; and (iii) any additional materials shared (e.g., email questions, presentations). We adopted a *qualitative content analysis* approach to explore the views, motivations and experiences of the project leads, and to address our more practical design

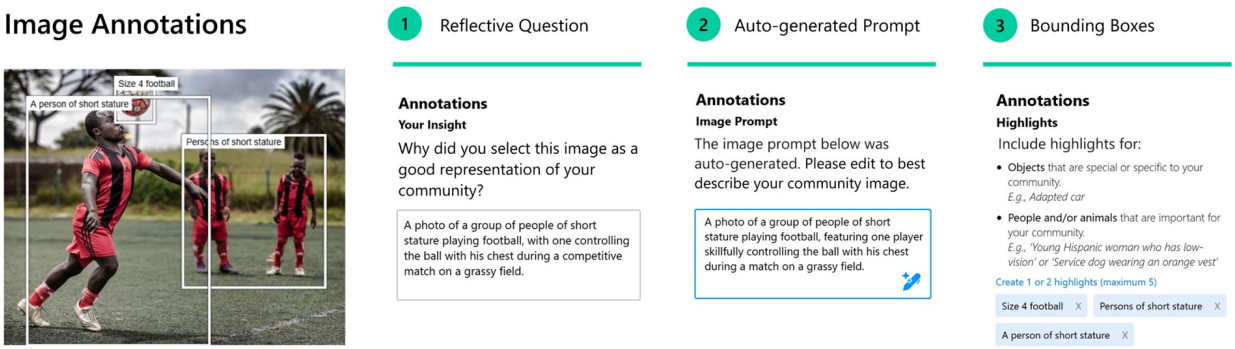


**Figure 3: Illustration of the three-step annotation process for each image of the Community Library.**

questions [37]. This method is well suited to our evolving process [37, 70]. It supports: for data collection and analysis to occur concurrently; the review of each study case (e.g., individual community) alongside the data as a whole [37]; as well as different content types [70] – from research notes and transcripts to system entries and rating results. We labeled our data with participant ID, engagement week, and data type (e.g., system entry, meeting). Our analysis involved: (i) data *immersion*, with "reflective notes" written by one or more researchers after each activity; (ii) data *coding*, whereby research findings were deductively categorized based on research phase, activity or question; while allowing themes to emerge inductively; and (iii) *interpretation*, through writing descriptive summaries with an interpretative narrative explaining how our judgments as researchers[2] were formed [37].

## 4 Findings

Organized around the three project phases, our findings show: (1) the individualized, dynamic, and negotiated nature of how each community defined representation and curated data supported by technology scaffolds; (2) challenges in aligning human insights with AI requirements; and (3) project leads' experiences of participating in AI data practices.

### 4.1 Defining Representation

To guide project leads in defining their community's representation, we designed 'structures' of the *Pinboard* to surface initial themes through imagery, and the *Magazine* to prioritize key themes and their representational variances. We also supported their planning of *Community Outreach* to integrate broader member perspectives.

[2]**Positionality**: Grounded in a constructionist epistemology, we view knowledge and meaning as actively and socially constructed; and acknowledge the researcher's interpretive role in identifying patterns through deep data engagement [18, 19]. All members of the research team are Western-based and employed at Microsoft at the time of the research, with three members having lived experience of disability. Working actively at the intersection of AI innovation, HCI and inclusion, we believe this is a pivotal time to address misrepresentation in AI and are committed to centering communities in AI development, ensuring their involvement is meaningful and beneficial. Beyond a focus on technical advances, we seek to improve inclusion and the lives of people with disabilities.

*4.1.1 Pinboard.* Each project lead began by uploading 10-12 images to their Pinboard, which they drew from available sources like their membership database, community socials or friends, and in two instances from online. This short activity produced nuanced accounts of what matters to a community, and how it is best articulated visually. For example, P1 selected an image of two girls with vision impairments (VI) in school uniforms, smiling as they each use a digital Braille device, with bulky Braille books covering the shelves in the background. Explaining the image's relevance in creating an inclusive learning environment and breaking barriers for VI learners, P1 states:

> "This image speaks volumes about access to inclusive education, especially for girls and transition to digital Braille for visually impaired learners. It represents the ongoing efforts to fully transition from manual Braille books to digital books and how this transition makes access to quality education a reality for girls with VI. This image further challenges the stereotype around educating girls and visually impaired learners in general [that they would not benefit from education]." (P1, wk3, system entry).

We made three additional observations of how project leads approached the representation definition process:

Firstly, we saw that all project leads tended to articulate their representation aspirations through image instances that would implicitly or explicitly challenge existing misconceptions and common stereotypes — aiming to ***counter 'negative' or 'limiting' narratives of their community in defining positive aspirations***. For instance, P2 included a photo of three footballers of short stature mid-game (Figure 3) to demonstrate their capability of playing demanding sports with appropriate modifications (e.g., smaller ball size). Explaining her rationale:

> "The image challenges assumptions that dwarfism limits physical capability or competitive spirit of athletes with dwarfism since the game of soccer is usually tied to height and speed; therefore proving that athleticism is defined by passion, not body size" (P2, wk3, system entry).

This definition of positive representation in contrast to negative examples highlights ***the importance of context — whether drawn from personal experiences or media history — in articulations of 'good' representation***.

Secondly, we observed challenges in balancing representations of the community's current, lived experiences and imagined future aspirations. For example, one Pinboard image prompted discussion about 'separation staring' – a behavior where average-height people 'gawk' or 'have some kind of reaction' (P3) to seeing a person with dwarfism that reflects current experiences. When asked whether AI could help reimagine future representations of their community – particularly by portraying individuals as more socially integrated to discourage behaviors like staring – P3 expressed openness to the idea, but also emphasized the importance of AI outputs that accurately reflect lived realities:

> "I think I would still feel validated if some pictures were created [with AI] that still have some level of this separation staring, because as much as we could paint rainbows and unicorns and sunshine of everything is perfect and people with dwarfism are accommodated for, are integrated and accepted, and are mixed in with other groups of people, there is still a sense of pride or ownership in where our community has come from in the history, that it hasn't always been that way. It's not yet perfect. I think I just imagined someone like me, someone from my community searching and saying like: Oh yeah, images will still come through that still have that accurate true to life piece." (P3, wk3, meeting)

A very practical challenge in envisioning a more inclusive future also lies in the scarcity of authentic imagery, particularly for activities that are yet to be realized, or rarely documented. P1 noted difficulties to find a single image online that didn't rely on stereotypes —such as the typical portrayal of a white man with sunglasses, cane, and guide dog — none of which he considered as a 'good reflection' of his community. This highlights how ***familiar visual norms can constrain representation***, and demonstrates how the Pinboard activity encouraged project leads to ***imagine aspirational representations that expand what is typically seen or considered possible today***.

Lastly, we observed how project leads began to actively define the boundaries of their communities. For example, P1 selected an image of a technician repairing Braille machines to signifying the 'labor involved' in ensuring inclusion of VI individuals. For his organization, the image: "reflects the value we place on access, maintenance and sustainability in supporting inclusive education" (P1, wk3, system entry), which extends definitions of community membership beyond PwD to include supporting roles, including sighted teachers and students. Similarly, the Global North dwarfism organization emphasized the inclusion of many average-height individuals (e.g., family members) in their membership. This illustrates ***the distinction between representing a community through an 'advocacy organization' versus a group of 'people with disabilities' for which boundaries of inclusion however can be difficult to define***.

*4.1.2 Magazine.* Across the three organizations, a range of higher-level themes were chosen to pursue in the Magazine activity: assistive technology (P1), independence (P1), ability (P2), diversity (P2, P3), (social) inclusion (P1, P2), empowerment (P1), pride/ joy and celebration (P3), learning/ education (P1, P2) and family (P3). Project leads then refined and prioritized their representation themes. For instance, for the VI organization, themes of 'education', 'assistive technology', and 'social inclusion' remained central, while 'empowerment' and 'independence' shifted towards a stronger focus on 'professional contributions to the economy' and 'sports'. At this stage, project leads felt five thematic feature pages were sufficient to express their core representation aspirations.

For the Magazine, all three organizations included 'work' as a theme, but each interpreted it differently: P3 focused on normalizing everyday work by showing people with dwarfism doing their jobs, working remotely, or multi-tasking. P2 emphasized professional achievements and showing people of short stature in leadership roles (e.g., a politician) – alongside more informal professions such as bartenders, or tea pickers. For the VI community, P1 showcased the broad accessibility of roles, which extend to the digital economy and creative industries, illustrated through images of a blind DJ, or a VI person working as photographer. This highlights that ***although organizations shared the importance of certain representation themes, each defined them in their own way***.

Interlinked with this, we observed geographical and cultural differences in representational priorities. The Global North dwarfism community emphasized themes of 'family', 'celebration' and 'friendship'; while the two Global South organizations featured 'education' as central to improving future opportunities for PwD, opening doors to work and income (P1), and enabling fuller participation in life (P2). For example, for P1, the theme of education for people with VI is closely entwined with, and 'powered by digital assistive devices'. Describing his representation aspirations, he emphasized the importance for images to not only correctly depict assistive technology (AT), but to also showcase the: (i) connections between humans and technology for accessing information; and (ii) interactions with other people. Consequently, almost all images P1 selected for education depicted important ATs for his community 'in-use' and across various learning contexts – such as a range of curricular activities, and different learning dynamics (e.g., a blind teacher passing their skills on to a VI student; a VI learner with sighted peers in a classroom). Critiquing how many images of ATs would not show a close-up of the person interacting with the device, he explained:

> "Your hands should be on the device, and this will be a good representation of how a person really calls the device and even the positioning of the device. I found that makes this a good image. Just show somebody who's in action using an actual assistive device, and also access to just to remove the negative perceptions around access to STEM subjects." (P1, wk4, meeting)

The organizations varied most in how they conceptualized 'diversity'. For the VI organization, P1 embedded diversity within other themes to highlight the spectrum of impairment – from low vision to full blindness – and differences across urban and rural contexts. In the Global South dwarfism community, P2 highlighted

the community's rich cultural heritage, spanning 42 tribes – expressed through unique languages, traditional clothing, and cultural artifacts. Lastly, P3, who defined 'diversity' as a main theme for their community, emphasized a wide range of: ages, races, dwarfism types, and family configurations. Here, P3 notes the challenge of representing ~400 different dwarfism types. While many share visual similarities, this necessitates decisions about balancing common and less prevalent body types. Project leads ***navigated making those choices by deliberating what meaningful variations to capture and what aspects to prioritize (for AI) in discussions with other community members***.

Finally, the Magazine prompted project leads to iterate or reframe what to include in images. For instance, P3 re-cropped an image of dwarfism entertainers on stage to include the audience – members of their community – to reduce risks of stereotypical portrayals (e.g., people with dwarfism being exploited for entertainment purposes), and instead to highlight a sense of community and shared enjoyment. This illustrates how the Magazine metaphor – that foregrounds how others may come to perceive a community through selected imagery – encouraged project leads to ***critically reflect on 'external' perceptions of their communities, prompting more intentional, nuanced choices in image selection and presentation***.

*4.1.3 Community Outreach.* Community feedback was central to shaping representation definitions. Each project lead independently mobilized their community and designed their own activities (see Table 2). Across the organizations, these engagements re-emphasized previous and surfaced new themes, including: portraying PwD in *professional roles* to foster positive narratives (P1, P2, P3); challenging misconceptions of PwD not having the right, or being able to raise a *family* (P1, P2); and ensuring *accurate depictions of bodies or devices* (P1, P2). For dwarfism, this meant showing the right body proportions (P2); for VI, community members emphasized realistic eye representation – avoiding distorted or enlarged depictions – and the correct depiction and handling of ATs (P1).

To engage members meaningfully in conversations about representation in AI images, project leads needed to clearly articulate the project's purpose and address concerns. To support in developing AI literacy, we provided project leads with various educational materials such as *presentation decks*, *webinars*, and a *progressively evolving FAQ sheet* that exemplified the project motivation and core AI concepts such as 'how AI generates images' or 'how image generation models learn'; and also responded to community-raised concerns such as 'fears of image use for bullying', or 'member's faces being recognizable in AI outputs'. These resources equipped project leads to more confidently explain core concepts and motivate participation. P3 expressed:

> "Even used the example of like the bird and the tree [from project's education resources] that like, you know, AI can learn patterns we don't want. And I really think that kind of made sense to a lot of people who have had a lot of questions or reservations about the project. " (P3, wk5, meeting)

Lastly, we learned how project leads ***navigated complex negotiations in defining 'good' representation with their community, balancing diverse perspectives with organizational goals and making choices where consensus cannot be achieved***. For example, during an in-person workshop that was attended by 'urban'-living VI individuals in Kenya, community members expressed concerns about the inclusion of portrayals of VI individuals in 'rural' contexts, which some feared could reinforce stereotypes. Simultaneously, they acknowledged the risk of marginalizing rural voices and disregarding their representational needs. To address this, P1 advocated a democratic resolution by consulting rural community members and suggesting a follow-up workshop to find compromise if tensions persisted:

> "(...) people who are in the rural area, they want to be represented in a way that reinforces negative stereotypes, and that might affect the long term goals, but also the other position is, if you and I use this in a light way, if you just impose what you feel is positive representation, it might also limit them in terms of their own identity, which we might differ. (. . . ) I think a broad compromise is much better than us agree saying this is the right way to do it." (P1, wk5, meeting)

## 4.2 Curating the Community Library

The Community Library offers a 'thematic structure' to help project leads refine initially defined higher-level representation themes (e.g., VI education) into specific sub-themes (e.g., blind teachers; primary students; non-curricular learning). See Appendix A.2, Figure 14 for an organization example. This structure guides the *curation of their community's 400-image dataset*; and serves as foundation for *image annotation*.

*4.2.1 Image Dataset Curation.* Project leads described the Community Library structure as helpful in planning and organizing their image collection, enabling them to track progress and define success. All organizations reported being 'very' satisfied (4/5) with their communities' representation in the Community Library in our closing survey, though they expressed interest in increasing image quality (P2) and diversity (e.g., more children and elderly (P2), family and groups (P1), rarer disability types (P3)). They curated their datasets via member donations (All), targeted image capture (All), and existing archives (P2, P3). These methods shaped their datasets: donations (P2, P3) yielded more individual portraits (P2, P3), while in-person efforts produced more group pictures (P1). Furthermore, (i) *data availability* via donations or existing archives; and (ii) *direct encounters with the lived experiences of PwD* prompted project leads to add new sub-themes. For example, after meeting VI farmers during rural visits, P1 introduced 'farming' as a sub-theme to highlight diverse economic roles, countering earlier concerns that rural portrayals might reinforce stereotypes:

> "The only iteration that you have to make is something we didn't consider in the beginning, the one on farming, because when we were starting, we assumed visually impaired people didn't farm, but quite the opposite. We found a lot of farmers, who are visually impaired, and we thought that's a good subcategory to add." (P1, wk9, meeting)

This illustrates how data curation enabled project leads to broaden and clarify their community and representation definition. Yet, for

| **Outreach details** | **Organization 1** (Vision Impairment, Global South) | **Organization 2** (Dwarfism, Global South) | **Organization 3** (Dwarfism, Global North) |
|---|---|---|---|
| *Recruitment Method* | Recruitment through member networks; members applied and were selected to ensure diverse representation | Facilitated by community leaders, WhatsApp outreach, and personal contacts | Internal recruitment of board members and committee chairs |
| *Activity Format & Attendee Profile* | Full-day, in-person workshop with 16 VI individuals | Consultation of 57 people of short stature via: (a) Individual online interviews (37 members); (b) Two in-person group interviews (1.5 hours each, 9 members/ per session) | In-person meeting with 5 board members; followed by online meeting with 18 board members and committee chairs |
| *Demographics* | Younger adults; urban (Nairobi-based); gender-balanced; in formal employment | Adults (21–61 years); majority women; regional diversity within Kenya | Young adults (professionals) to older adults; geographically diverse within the USA; the vast majority have dwarfism |
| *Absent Groups* | Rural members; younger children and students | Elderly members of the community | LGBTQ sub-community; inclusion director; young adults coordinator |
| *Language Used* | Mainly English, with Swahili expressions for clarity | Swahili predominantly, with some English | English |
| *'Magazine' Theme Iteration* | High | Medium | Low |

**Table 2: This table outlines the outreach approach taken by each of the three disability organizations, detailing their recruitment method, format of activities, attendee profiles and demographics, inclusion of languages and extent to which the community engagement influenced iterations of the Magazine themes.**

this to work, project leads needed to ***continuously align their evolving representation goals with AI requirements for a 'balanced' dataset*** – an effort they described as both time-consuming and cognitively demanding. Although the Community Library structure proposed five themes with four sub-themes (~20 images each), we were careful to not over-enforce but to balance this with communities' representation goals and feasibility constraints. For example, the Global North dwarfism community chose three rather than five main themes for their library, resulting in 23-46 images in each of four sub-themes. Where more images fit a sub-theme, project leads decided to add or split sub-themes. Where image sourcing was difficult, we did not enforce any minimum numbers, meaning that the final dataset of the VI community had various sub-themes with fewer images (e.g., digital economy (7), farming (13), competition or tournaments (13)). Furthermore, sub-themes like 'graduation' needed fewer images to portray, while themes of 'work' – broken down into 'formal', 'informal' or 'creative' industries – required more images to reflect the diversity of potential professions and across settings. This underscores that ***themes are not equal and highlights limitations in this approach, where some themes may need more images or finer sub-theme distinctions***.

Finally, organizations navigated practical constraints in capturing desired, diverse representations – especially securing AI-use 'consent' for images of 'groups' or 'children'. Recognizing however the importance of including families, friends or colleagues for representing the lived experiences of PwD, project leads adopted community-led strategies. For example, P1 consulted members on consent protocols for images that showed more than one individual (e.g., should the person owning the image be able to give consent on others' behalf?); jointly they agreed to exclude images that showed strangers, and to seek consent from known others (e.g., family). To include children, P1 leveraged existing contacts with schools for classroom captures and securing guardian approvals; whereas P3 bypassed guardian consent altogether by having adult members with dwarfism donate their childhood images. These examples show how ***communities exercised agency in shaping consent strategies and leveraged their networks to meet their inclusion goals***.

*4.2.2 Image annotations.* For image annotations, project leads appreciated the reflective 'why' question as a clear, non-technical way to describe image relevance. As exemplified in Figure 4, this approach produced rich descriptions that foreground the representational relevance of an image to external audiences, or AI models; as well as serving as the basis for image-level prompt generation for AI evaluation. Within this configuration, project leads were also able to ***specify preferred terminology for their community, reflecting their cultural values and ensuring accurate, respectful and inclusive language***. For example, P2 favored the term 'people of short stature', whilst P3 preferred 'people with dwarfism'. P3 described how being given the agency to define this terminology felt empowering:

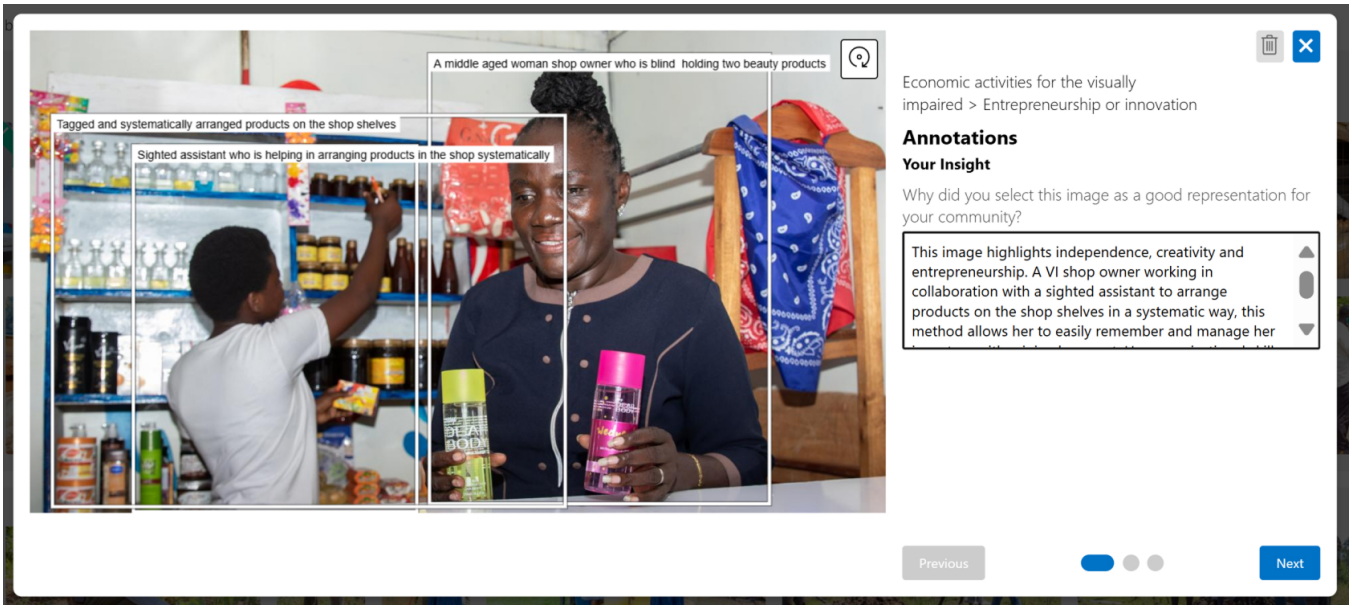


**Figure 4: Example how project lead P1 described his rationale for selecting an image as good representation for their community. Later added, labeled bounding box annotations further illustrate added detail to the characteristics of the people portrayed. Full annotation text: *"This image highlights independence, creativity and entrepreneurship. A VI shop owner working in collaboration with a sighted assistant to arrange products on the shop shelves in a systematic way, this method allows her to easily remember and manage her inventory with minimal support. Her organizational skills, memory, and ability to adapt quickly make it easy to run the business effectively".***

> "(…) we are very used to things being decided for us. Someone else thinking that they know best, someone else thinking that their degrees or their experience gives them the permission to decide things on behalf of disabled people. So just the chance to get to participate in something like this where we are being asked to provide the data, but we're also being asked how we want to be represented. You all allowed us to choose the language we wanted and how we're represented, and all of that just empowers us to get to do that in other spaces in the world. (…) just have any kind of say that comes directly from us as the experts was truly a life changing opportunity and I think that's going to take even more time to really settle in how big of an impact that has had". (P3, wk14, webinar)

For soliciting labeled bounding boxes, however, we had to iterate on our designs. Initially, when we asked project leads to highlight image elements they considered important for their community, they were unsure what to highlight that was of relevance for AI, which led to over-annotation of general items (e.g., a wall clock) and inconsistencies across images. To address this, we limited annotations to two categories of community relevant 'objects' and 'people/animals', and capped each image at five bounding boxes to encourage meaningful selections (see Appendix, Figure 13). This created ***greater alignment of the annotation task with their community representation, which improved the relevance of their annotations***. Figure 4 further shows how bounding box labels were often utilized to add diversity dimensions to 'people' such as gender, age, level of VI, and actions. However, there were still some inconsistencies whereby relevant aspects in an image could be overlooked (e.g., labeling a footstool in some images, but not others), and a lesser nuancing of 'object' descriptions (e.g., Appendix, Figure 13 has the 'white cane' labeled, but not that it is 'folded'). Notably, P3 developed her own, systematic template for creating bounding boxes, which was not only motivated by the need for data consistency; but also brought a 'relational' sense of accountability to their community regarding how annotations were made:

> "And then the highlights [term used for labeled bounding boxes]. Again, I've kind of kept a very similar structure of like any person I'm highlighting. I do like age, gender, culture or race type of dwarfism and then what they're doing.(...) That's also a question I received from members when donating photos (...) can you tell me how is my photo going to be talked about, or what language is going to be used. I had many people ask me like, is [the research team] deciding (...) so it was cool to be able to tell them like, no, we got to decide. I got to decide on behalf of the dwarfism community (...) So being able to kind of give that answer, (...) felt reassuring." (P3, wk17, meeting)

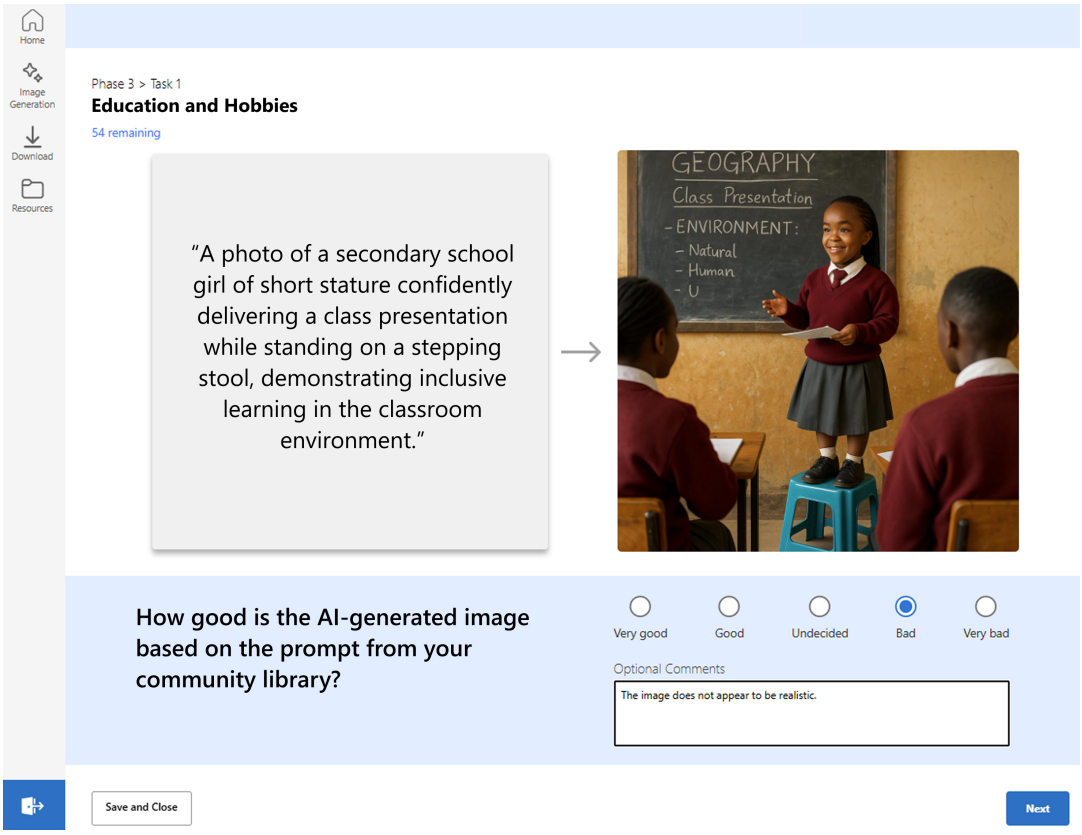


**Figure 5: Example illustrating the AI evaluation rating task that shows the prompt that was generated based on a real image of P2's Community Library vis-a-vis the generated AI image (from one of three image generation models), and its rating as 'bad' with optional commentary.**

## 4.3 Evaluating AI

For the AI evaluation task, we generated images from Community Library prompts and asked project leads to assess their 'goodness' instead of focusing on representational harms or AI errors. Project leads expressed their community-specific representation preferences across the full spectrum of the 5-point rating scale, designed to capture holistic judgments across ~300 AI images as input to building AI evaluator models. Through optional comments and discussions in research meetings, project leads explained both the shortcomings and strengths of AI generated images.

We learned how 'very good' representation evaluations were not only assessed based on correctness, but often associated with how *realistic or natural* an image depicted people – showing the individual(s) via appropriate body proportions or attire; and as engaged in *authentic* interactions within *believable* settings (e.g., a realistic-looking marketplace). Variances between good or very good images were often attributed to how *respectful* the depiction felt; the *mood* captured (e.g., is the person showing enjoyment); and the fit of smaller details with annotators *expectations* (e.g., there should be more tools if it is to show a repair setting). Furthermore, scores were lowered if images *misaligned with the prompt*, for example by missing or mis-representing key aspects (e.g., no evidence of the person being a 'teacher'). Images were rated as 'bad' or 'very bad', where the *disability was either not recognizable or mis-portrayed*. This included few cases marked as infantilization and renderings of average-height people for the dwarfism communities. Scores were also deducted for *inaccurate rendering of assistive devices* and occasionally *hallucinated artifacts* (e.g., disproportionate renderings of hands or feet). These qualitative insights surface how evaluations are not just based on objective, physical criteria (e.g., missing an instance given the prompt), but ***reflect annotators interpretations and perceptions of the 'authenticity', 'respectfulness' and 'appropriateness' of an image with regards to their specific community and the context shown***.

Furthermore, we observed how the holistic rating score often reflected a combination of different evaluation criteria. For example, in her assessment of the image in Figure 5, P2 explained that although the demonstration of the blackboard is 'good', the image does not depict the person of short stature 'realistically' as she resembles more 'a child wearing a uniform surrounded by average height students'. She also explains insufficiency with the rendered stepping stool:

> "(…) on this one, you can easily and quickly drop down and get some injuries. The kind of stepping stool that usually help us to climb on something that you want probably to make your height more average

> is foldable and it has another step. So you cannot easily like fall down." (P2, wk13, meeting)

This shows how ***project leads applied their specific domain expertise to identify subtle nuances and make careful judgments about the images***, by paying close attention to details of the stepping stool that did not align with their own experiences or understandings of what is considered appropriate or logical in this situation. This demonstrates a broader challenge in image generation models to interpret the functional, embodied realities of assistive technologies (e.g., is this the right stepping stool for this community) beyond often more simplistic assessments of objective features of the physical world (e.g., whether a stepping stool was generated or not). Appendix A.3 includes additional AI evaluation examples.

## 5 Discussion: Bringing Communities Meaingfully into AI Data Practices

In this paper, we argue that to advance disability representation in image generation models requires a reconfiguration of how disability communities are meaningfully brought into AI data practices that underpin model development and evaluation. Building on calls for more inclusive, data-centric AI work (e.g., [11, 31, 53, 96, 107]), we present our approach to centering disability communities within important AI data practices through a technology-supported, community-led definition and curation process – scaffolded via the Community Library Creator. We conclude by (1) discussing derived insights on how communities defined representation through this process; (2) examining the value and practical challenges of aligning human insights with technical AI requirements; and (3) reflecting on human-centered AI approaches that empower communities to contribute their perspectives and actively shape AI.

### 5.1 Technology-Supported, Community-Led Representation Definition

Responding to recent calls for more community-based and participatory methods in defining and evaluating representation for AI [10, 68, 89, 90], we described our approach to providing disability advocacy organizations with technology scaffolds and facilitating opportunities for community dialogue, which enabled the production of nuanced, contextual accounts of what 'good' representation means for each community, and its embedding within data for AI.

Echoing recent research by Qadri et al. [89], our findings show how each community actively negotiated their own representation goals – shaped by their social, geographical and cultural contexts, and in conversation with community members and image materials. Each community defined representation in their own, distinct ways, which is reflected in diversity across their representation themes that spanned family, celebrations, or sports as well as within themes such as those of 'work' and 'diversity'. Furthermore, our findings surface rich nuances in communities' definitions. This was best illustrated in detailed accounts of how assistive technology (ATs) should be depicted as being 'handled' by a person, 'used within socially inclusive settings' and 'for STEM learning'. All this highlights the *contextualized, interpretative* nature of definitions by each community that clearly extend the scope of relevant criteria to consider beyond the accuracy of observable aspects of the physical world (e.g., is the AT in itself correctly rendered). It suggests the need to develop methods for defining and evaluating representation in image generation AI that are capable of reflecting the diversity and plurality found across different communities. In line with arguments by [68, 89, 90], such ***approaches should recognize the inherently subjective meanings and perspectives that shape social concepts like representation, thereby moving beyond singular or objective standards towards embracing the richness of community-specific interpretations***.

Furthermore, our findings surface how social worlds and their ideal representations are not static, but can evolve over time in response to changing contexts and aspirations (cf. [89]). This was most evident in the example of 'separation staring' in dwarfism, which surfaced challenges in balancing a communities' current, lived experiences with more desirable future aspirations. Our work also demonstrates how defining representation is fundamentally a *discursive* process, shaped: (i) by negotiation and dialogue within communities (e.g., deliberations of urban vs. rural depictions); (ii) in response to broader societal narratives (e.g., desires to challenge existing stereotypes and common misconceptions); and (iii) through material engagements with the images. For example, community contributions such as donating images and direct engagements with member's lived experiences (e.g., discovery of visually impaired farmers) expanded both the scope and inclusion of representation in Community Libraries. In dwarfism, we describe how image manipulation, such as cropping, is utilized to reduce risks of stereotypical portrayals (e.g., exploitation of people with dwarfism for entertainment purposes). Lastly, further nuance is added through image descriptions (Figure 4), image highlights (e.g., adding diversity dimensions), and qualitative comments given with ratings on generated AI images. These findings show that ***material engagement with images—through data curation, manipulation, and annotation offers an additional pathway for communities to refine their representation definition and embed these nuanced understandings directly within data***.

### 5.2 Intersecting Human Insights with Technical AI Requirements

Our findings show how reflection scaffolds and data structures provided through the Community Library Creator – such as Pinboard, Magazine, the Community Library, image annotations, and AI image ratings – served as a productive bridge for project leads to articulate their communities' representation goals in data. However, we acknowledge that by choosing these structures, we stirred the direction of how communities approached defining representation.

One key structure for organizing community inputs is the Community Library. Through its design, it imposes a 'thematic structure' and requirements for a 'balance' in data across thematic groups, which could inadvertently risk pigeonholing communities' in the ways they may wish to express themselves or categorize information. Our findings uncover tensions for project leads to achieve a fully 'balanced' dataset, noting: (i) data availability constraints; (ii) difficulties to match the scale of diversity (e.g., 42 different tribes, >400 dwarfism types, variety in professions) into a roughly equal number of images across thematic (sub)themes; as well as (iii) inequality across themes, where some may require finer sub-theme

distinctions than others. Our approach also needed to remain flexible and open to aligning data balance with what is desired and possible for communities to source through their network, which necessitated ongoing negotiations and thematic re-clustering. While this process can be effortful, we found the Community Library to have served as a useful reference for organizing project leads thinking and outreach activities by reducing uncertainties about how much data and what type of data to collect for a target dataset (here: 400 images), whilst the contents themselves help preserve the richness and authenticity of community perspectives. As such, ***the Community Library structure provides a useful boundary for supporting communities in prioritizing their representation goals whilst making data requirements for AI more concrete***.

Another key structure is 'data annotation'. Here, our approach varies from many traditional protocols in two ways: Firstly, we took a community-centric approach to the creation of 'image descriptions' and the 'prompts' we derived from it, by focusing less on 'what' an image entails and instead asking 'why' the image meaningfully reflects the community. Through this we exemplify how we can ***bring a greater emphasis on human 'values' into data annotation practices***. Secondly, our annotation process did not rely on pre-defined protocols, commonly carried out by anonymous, crowd-sourced data labellers (e.g., [53, 85]) that often lack in-domain expertise [51, 52]. Instead, we worked with project leads as in-domain, community advocates. However, for the process of creating image 'highlights', our findings also reveal difficulties in enabling effective annotations despite provisions of clear instructions and educational materials. Project leads initially struggled to identify relevant bounding boxes and specify label details that would be of relevance for AI, leading to over-annotation of general items, and inconsistencies. Variation in how annotators subjectively interpret and follow concrete definitions has also been noted in other image generation data annotation work [52]. However, we saw that by re-shaping our bounding box label instructions to focus on people and objects specific to the community, we created greater alignment of the annotation task with their community's representation. This ***closer coupling of the annotation task around the notion of 'good' representation improved the relevance of annotations over more generic labels***.

Lastly, in our 'AI evaluation', we observed complexity in most image assessments that were often based on a combination of multiple evaluation criteria and annotator interpretations and perceptions of AI generated images – assessing authenticity, respectfulness and appropriateness with regards to representations of their specific community and context – which clearly extends beyond evaluations of objective features of the physical world. This echoes recent works [10, 68, 89, 90] that argue against treatments of representation as a static concept that can be evaluated objectively. Instead, we ***advocate for new AI evaluation approaches that better capture the contextual, interpretative nature of a particular communities' representation***. In our work, we recorded project leads perspectives and preferences in qualitative comments as well as holistic image goodness 'rating scores' – aiming to better capture their representation preferences and their integration into the development of new AI measurements (cf. [77, 81, 89–91]). In creating new, more community-centric measurements, a fundamental challenge remains in ***balancing people's ability and the effort required to capture subtle distinctions and deeper interpretations for each image*** – especially for specific community knowledge, values, and norms that are often harder for people to articulate or quantify [68, 89] — ***with the needs for easier-to-scale, holistic measurement tools***.

## 5.3 Human-Centred AI Data Work

*5.3.1 Shifting from Identifying 'Harm' or 'Errors' to Cultivating Meaningful Representation.* Our approach to defining and evaluating 'good' representation differs from existing research that predominantly focuses on understanding, defining or evaluating representational 'harms' [24, 28, 38, 44, 67, 77, 91, 116] and often uses poor AI-generated images to elicit feedback from marginalized communities (e.g., [65, 77, 79]). Other trends in AI evaluation include red-teaming, whereby image generation models or systems are deliberately tested with adversarial prompts to uncover their 'vulnerabilities' and 'bias'; surfacing offensive or harmful content for purposes of error detection, or removal [46]. However, researchers caution that involving marginalized communities in elaborating on their marginalization for red-teaming, can become transactional, extractive and exploitative (cf. [29, 46]); and warn of the emotional and mental health costs of this type of human 'labor'. In our work, we therefore specifically chose to ***minimize community exposure to overly negative AI outputs***, for example, by filtering out clearly offensive images for our AI evaluation task.

Rather than focusing narrowly on identifying or correcting 'errors' – whilst perhaps a cognitively easier to define task – we adopted a positive approach throughout all our system scaffolds: asking for uploads of 'good' images or prompting reflection on why an image was a 'good' representation. This strategy ***centers the community's lived experiences and aspirations for more meaningful representation rather than requiring them to deeply engage with the limitations and failings of image models***.

However, our findings showed how a deliberate focus on 'good' representation, did not mean that negative examples would not surface. All project leads articulated common problems with stereotyping and lived experiences of disability used to anchor expressions of 'better' representation and deepening their definition. Importantly, such articulations of negative narratives were not imposed upon project leads, but evolved naturally, leaving them in control as to when and how they would engage with negative instances. As we continue to imagine the best ways to support human-centred AI data practices, ***we propose not to eradicate negative representations, but rather consider what kind of choice (or control) people have in engaging with negative material***.

Finally, our emphasis on defining 'good' representation was experienced positively by project leads, who described their enjoyment collaborating with and learning from their community members – especially during workshops and field trips. They expressed pride in the resulting image collection, and intentions to continue building on these efforts. The inherent social nature of data production and the enjoyment from working 'together' towards a shared goal, echoes findings from other community-oriented AI dataset specification and curation work [32, 99, 107]. Such experiences highlight the value of taking a human-centered approach to AI data practices

that ***focuses on cultivating positive representation; and prioritizes the meaning that communities derive from participating over the effort or time required*** – as might be considered in more transactional views.

*5.3.2 Community-Defined Boundaries & Infrastructures for Outreach and Deliberation.* Taking a community-centric approach to AI data practices gives us one perspective on 'who' gets to be involved, and 'whose values and norms' get to shape AI pipelines. Researchers have previously problematized anonymous crowdsourced data annotation approaches [31, 53, 65] as being unspecific about who's worldviews and beliefs become manifest in the data [11, 17, 51, 78, 90, 91]. In more community-centric data collection or AI evaluation work, active efforts are made by researchers to solicit broad participation (e.g., across geographical regions [52, 68, 91] and through local research partners [89]), or allow community members to self-select their participation, albeit with some limitations (e.g., adults, higher-education levels or topic expertise, English-language ability) [68, 78, 99]. In this work, we chose to recruit and ***specifically define communities as a bounded set of people represented by the membership of a disability advocacy organization***. Rather than treating PwD as a singular or universal identity – often reinforced by homogenizing labels such as 'dwarfism' – we acknowledge the distinct social identities and intersectionalities within each community (cf. [53, 90]). In our research, this was most evident in differences in the representation priorities, disability language, and scope of diversity of the two dwarfism communities.

We found that engaging with disability advocacy organizations as proxies for bounded communities provides a practical and realistic approach to involving non-AI experts in AI data practices that respects each communities' own ways of organizing themselves, and their culturally embedded power dynamics. Our findings show how each organization mobilized their community differently (Table 2), and how this shaped participatory inclusion and deliberation. Project leads spoke local languages and had intimate knowledge of various sub-groups within their community to identify individuals or regions who may be underserved. This played a key role in accessing and connecting with more marginalized, harder-to-reach populations (cf. [53]) such as blind farmers in rural areas or school-aged children, who otherwise may not have known about, or understood why they should participate. Advocacy organizations also already have established processes to support communication and decision-making, enabling them to address tensions and differing perspectives across their membership. We saw how they deliberated issues of consent and negotiated the boundaries of their community with their members. While no form of participation or negotiated outcome will be perfectly inclusive, ***this approach ensures that decisions about whose voices are represented in the data remain with the communities themselves, rather than external AI developers***.

## 5.4 Limitations & Future Work

This work has several limitations. It is grounded in engagements with three disability organizations to enable in-depth, iterative dialogue about how project leads navigated their involvement in AI data practices and community engagement. Future work is needed to examine how these insights generalize across a broader diversity of potential organizations. As such, there is a need for evaluative studies with a wider range of communities to assess how easily communities could navigate the platform and process independently.

Our work foregrounds defining 'good' representation for AI and adopts a model agnostic approach to data curation and annotation. However, different image generation models exhibit distinct limitations and failure modes. Greater awareness of model specific errors could guide communities to curate data that explicitly targets these shortcomings. As a result, while our approach seeks to improve overall representation quality, it remains limited in its ability to eliminate model-specific errors. A similar generalizability challenge applies to the future development of community-specific evaluator models, whose training data is constraint to the three image generation models used in our AI rating task.

Open questions also remain about how best to use the datasets in a rapidly evolving model landscape. Through the Community Library, we support the production of multiple forms of *AI interpretable data* including: a 400-image dataset; image descriptions; structured bounding-box annotations with text labels; a thematic hierarchy of representation themes and sub-themes to support prompt creation; image-level prompts; and ratings of AI image–prompt pairs. While prior work suggests that small, high quality datasets can support *post*-training alignment and adaptation [86, 122]; more data intensive *pre*-training settings raise challenges around data scaling (without compromising authenticity). Future work is needed still to demonstrate how best to leverage the Community Library datasets to improve AI models or systems.

## 6 Conclusion

As AI media generation becomes increasingly prevalent, there is a critical opportunity to ensure PwD are accurately represented. Recognizing that improving image generation requires more than just data, we placed communities and their advocacy organizations at the center of decisions about their own representation. To lower barriers for non-AI expert communities to embed their values and language in AI data practices, we worked closely with three disability advocacy organizations to develop the Community Library Creator. This prototype provided various design scaffolds – alongside broader community engagements – to support with important AI data practices of: (i) concept definition; (ii) community-led data curation and structuring into AI interpretable formats; and (iii) preference elicitation in AI image evaluation. We discussed how communities defined good representation through this process; the value and practical challenges of aligning human insights with AI requirements; and human-centered AI approaches for empowering communities to share their perspectives – paving the way towards more inclusive and equitable generative AI.

## Acknowledgments

We gratefully acknowledge the community members whose contributions were central to shaping each organization's representation definitions and curating the datasets. We also sincerely thank our partner organizations and their dedicated staff and wider support networks for their leadership, participant mobilization, and careful stewardship, which ensured that the datasets authentically reflect the lived experiences and voices of their communities.

## References


[1] British Academy and the Royal Society. 2017. Data management and use: governance in the 21st century.

[2] Rudaiba Adnin and Maitraye Das. 2024. "I look at it as the king of knowledge": How Blind People Use and Understand Generative AI Tools. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 64, 14 pages. doi:10.1145/3663548.3675631

[3] Rahaf Alharbi, Pa Lor, Jaylin Herskovitz, Sarita Schoenebeck, and Robin N. Brewer. 2024. Misfitting With AI: How Blind People Verify and Contest AI Errors. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 61, 17 pages. doi:10.1145/3663548.3675659

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[5] Anonymous. 2022. *Seen on screen: The importance of disability representation*. Available from https://www.nielsen.com/[Accessed30-08-2025].

[6] Anonymous. [n. d.]. Disabled and Here Collection. Available from https://affectttheverb.com/collection/ [Accessed 09-09-2025].

[7] Abhipsa Basu, R Venkatesh Babu, and Danish Pruthi. 2023. Inspecting the geographical representativeness of images from text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5136–5147.

[8] Eric PS Baumer. 2017. Toward human-centered algorithm design. *Big Data & Society* 4, 2 (2017), 2053951717718854.

[9] Cynthia L Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P Bigham, Anhong Guo, and Alexandra To. 2021. "It's complicated": Negotiating accessibility and (mis) representation in image descriptions of race, gender, and disability. In *Proceedings of the 2021 chi conference on human factors in computing systems*. 1–19.

[10] Cynthia L Bennett, Shaun K Kane, and Christina N Harrington. 2025. Toward Community-Led Evaluations of Text-to-Image AI Representations of Disability, Health, and Accessibility. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 256–270.

[11] Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. 2024. STELA: a community-centred approach to norm elicitation for AI alignment. *Scientific Reports* 14, 1 (2024), 6616.

[12] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1493–1504. doi:10.1145/3593013.3594095

[13] Abeba Birhane, Sepehr Dehdashtian, Vinay Prabhu, and Vishnu Boddeti. 2024. The dark side of dataset scaling: Evaluating racial classification in multimodal models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1229–1244.

[14] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1536–1546.

[15] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[16] Elizabeth Bondi, Lily Xu, Diana Acosta-Navas, and Jackson A Killian. 2021. Envisioning communities: a participatory approach towards AI for social good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 425–436.

[17] Geoffrey C Bowker and Susan Leigh Star. 2000. *Sorting things out: Classification and its consequences*. MIT press.

[18] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology* 18, 3 (2021), 328–352.

[19] David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & quantity* 56, 3 (2022), 1391–1412.

[20] Stephanie Russo Carroll, Desi Rodriguez-Lonebear, and Andrew Martinez. 2019. Indigenous data governance: strategies from United States native nations. *Data science journal* 18 (2019), 31.

[21] Marc Cheong, Ehsan Abedin, Marinus Ferreira, Ritsaart Reimann, Shalom Chalson, Pamela Robinson, Joanne Byrne, Leah Ruppanner, Mark Alfano, and Colin Klein. 2024. Investigating Gender and Racial Biases in DALL-E Mini Images. *ACM J. Responsib. Comput.* 1, 2, Article 13 (June 2024), 20 pages. doi:10.1145/3649883

[22] Arnavi Chheda-Kothary, Ritesh Kanchi, Chris Sanders, Kevin Xiao, Aditya Sengupta, Melanie Kneitmix, Jacob O. Wobbrock, and Jon E. Froehlich. 2025. ArtInsight: Enabling AI-Powered Artwork Engagement for Mixed Visual-Ability Families. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 190–210. doi:10.1145/3708359.3712082

[23] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

[24] Jennifer Chien and David Danks. 2024. Beyond behaviorist representational harms: A plan for measurement and mitigation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 933–946.

[25] Madiha Zahrah Choksi, Sharon Heung, Reeda Shimaz Huda, Rita Faia Marques, Anja Thieme, and Cecily Morrison. [n. d.]. Terms of Care: Designing Participatory Data Governance for Disability Communities. ([n. d.]). [Manuscript in preparation].

[26] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[27] Ned Cooper and Alexandra Zafiroglu. 2024. From Fitting Participation to Forging Relationships: The Art of Participatory ML. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 746, 9 pages. doi:10.1145/3613904.3642775

[28] Emily Corvi, Hannah Washington, Stefanie Reed, Chad Atalla, Alexandra Chouldechova, P Alex Dow, Jean Garcia-Gathright, Nicholas Pangakis, Emily Sheng, Dan Vann, et al. 2025. Taxonomizing Representational Harms using Speech Act Theory. *arXiv preprint arXiv:2504.00928* (2025).

[29] Samantha Dalal, Siobhan Mackenzie Hall, and Nari Johnson. 2024. Provocation: Who benefits from" inclusion" in Generative AI? *arXiv preprint arXiv:2411.09102* (2024).

[30] Sylvie Delacroix and Neil D Lawrence. 2019. Bottom-up data trusts: Disturbing the 'one size fits all'approach to data governance. *International data privacy law* 9, 4 (2019), 236–252.

[31] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. 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* (Boston, MA, USA) *(EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 37, 23 pages. doi:10.1145/3617694.3623261

[32] Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. 2023. ASL citizen: a community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems* 36 (2023), 76893–76907.

[33] Emory James Edwards, Cella Monet Sum, and Stacy M. Branham. 2020. Three Tensions Between Personas and Complex Disability Identities. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3334480.3382931

[34] Katie Ellis, Gerard Goggin, Beth A. Haller, and Rosemary Curtis (Eds.). 2020. *The Routledge Companion to Disability and Media*. Routledge, New York. https://doi.org/10.4324/9781315716008

[35] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

[36] Anjalie Field, Amanda Coston, Nupoor Gandhi, Alexandra Chouldechova, Emily Putnam-Hornstein, David Steier, and Yulia Tsvetkov. 2023. Examining risks of racial biases in NLP tools for child protective services. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1479–1492. doi:10.1145/3593013.3594094

[37] Jane Forman and Laura Damschroder. 2007. Qualitative content analysis. In *Empirical methods for bioethics: A primer*. Emerald Group Publishing Limited, 39–62.

[38] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Remi Denton, and Robin Brewer. 2023. " I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 205–216.

[39] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Remi Denton, and Robin Brewer. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 205–216. doi:10.1145/3593013.3593989

[40] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2023. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems* 36 (2023), 27092–27112.

[41] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (April 2018). doi:10.1073/pnas.1720347115

[42] Sanjana Gautam, Pranav Narayanan Venkit, and Sourojit Ghosh. 2024. From melting pots to misrepresentations: Exploring harms in generative ai. *arXiv preprint arXiv:2403.10776* (2024).

[43] Kathrin Gerling, Arno Depoortere, Jeroen Wauters, Katta Spiel, Dmitry Alexandrovsky, Marina Danckaerts, Dieter Baeyens, and Saskia Van der Oord. 2024. Representation of Invisible Disability: Exploring the Lived Experience of Teenagers with ADHD to Inform Game Design. *ACM Trans. Comput.-Hum. Interact.* 31, 5, Article 58 (Nov. 2024), 26 pages. doi:10.1145/3685276

[44] Sourojit Ghosh, Nina Lutz, and Aylin Caliskan. 2024. "I Don't See Myself Represented Here at All": User Experiences of Stable Diffusion Outputs Containing Representational Harms across Gender Identities and Nationalities. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society*, Vol. 7. 463–475.

[45] Tarleton Gillespie. 2024. Generative AI and the politics of visibility. *Big Data & Society* 11, 2 (2024), 20539517241252131.

[46] Tarleton Gillespie, Ryland Shaw, Mary L Gray, and Jina Suh. 2024. AI red-teaming is a sociotechnical challenge: on values, labor, and harms. *arXiv preprint arXiv:2412.09751* (2024).

[47] Kate Glazko, Yusuf Mohammed, Ben Kosa, Venkatesh Potluri, and Jennifer Mankoff. 2024. Identifying and Improving Disability Bias in GPT-Based Resume Screening. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. ACM, 687–700. doi:10.1145/3630106.3658933

[48] Kate S Glazko, Momona Yamagami, Aashaka Desai, Kelly Avery Mack, Venkatesh Potluri, Xuhai Xu, and Jennifer Mankoff. 2023. An Autoethnographic Case Study of Generative Artificial Intelligence's Utility for Accessibility. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 99, 8 pages. doi:10.1145/3597638.3614548

[49] Tina Goethals, Dimitri Mortelmans, Hilde Van den Bulck, Willem Van den Heurck, and Geert Van Hove and. 2022. I am not your metaphor: frames and counter-frames in the representation of disability. *Disability & Society* 37, 5 (2022), 746–764. arXiv:https://doi.org/10.1080/09687599.2020.1836478 doi:10.1080/09687599.2020.1836478

[50] Google DeepMind. 2025. Imagen4. https://storage.googleapis.com/deepmind-media/Model-Cards/Imagen-4-Model-Card.pdf. Accessed: 2025-09-03.

[51] Nitesh Goyal, Ian D Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28.

[52] Melissa Hall, Samuel J Bell, Candace Ross, Adina Williams, Michal Drozdzal, and Adriana Romero Soriano. 2024. Towards geographic inclusion in the evaluation of text-to-image models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 585–601.

[53] Siobhan Mackenzie Hall, Samantha Dalal, Raesetje Sefala, Foutse Yuehgoh, Aisha Alaagib, Imane Hamzaoui, Shu Ishida, Jabez Magomere, Lauren Crais, Aya Salama, et al. 2025. The Human Labour of Data Work: Capturing Cultural Diversity through World Wide Dishes. *arXiv preprint arXiv:2502.05961* (2025).

[54] Brienna Herold, James Waller, and Raja Kushalnagar. 2022. Applying the Stereotype Content Model to assess disability bias in popular pre-trained NLP models underlying AI-based assistive technologies. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, Sarah Ebling, Emily Prud'hommeaux, and Preethi Vaidyanathan (Eds.). Association for Computational Linguistics, Dublin, Ireland, 58–65. doi:10.18653/v1/2022.slpat-1.8

[55] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2021).

[56] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[57] Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. 2024. 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*. 1–17.

[58] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20406–20417.

[59] Maui Hudson, Stephanie Russo Carroll, Jane Anderson, Darrah Blackwater, Felina M Cordova-Marks, Jewel Cummins, Dominique David-Chavez, Adam Fernandez, Ibrahim Garba, Danielle Hiraldo, et al. 2023. Indigenous Peoples' rights in data: a contribution toward Indigenous Research Sovereignty. *Frontiers in Research Metrics and Analytics* 8 (2023), 1173805.

[60] Shuxu Huffman, Si Chen, Kelly Avery Mack, Haotian Su, Qi Wang, and Raja Kushalnagar. 2025. "We do use it, but not how hearing people think": How the Deaf and Hard of Hearing Community Uses Large Language Model Tools. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 33, 9 pages. doi:10.1145/3706599.3719785

[61] Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. GenAssist: Making Image Generation Accessible. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 38, 17 pages. doi:10.1145/3586183.3606735

[62] Ben Hutchinson, Vinodkumar Prabhakaran, Remi Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Unintended machine learning biases as social barriers for persons with disabilitiess. *SIGACCESS Access. Comput.* 125, Article 9 (March 2020), 1 pages. doi:10.1145/3386296.3386305

[63] Ada Lovelace Institute. 2021. Participatory Data Stewardship: A Framework for Involving People in the Use of Data. *Report* (2021).

[64] Open Data Institute. 2020. Data Trusts in 2020. Available from https://theodi.org/news-and-events/blog/data-trusts-in-2020/.[Accessed17-04-2024].

[65] Nari Johnson, Hamna ., Deepthi Sudharsan, Theo Holroyd, Samantha Dalal, Siobhan Mackenzie Hall, Jennifer Wortman Vaughan, Daniela Massiceti, and Cecily Morrison. 2025. Position: To Make Text-to-Image Models that Work for Marginalized Communities, We Need New Measurement Practices for the Long Tail. (July 2025). https://www.microsoft.com/en-us/research/wp-content/uploads/2025/07/longtail.pdf.

[66] Samia Kabir, Lixiang Li, and Tianyi Zhang. 2024. STILE: Exploring and Debugging Social Biases in Pre-trained Text Representations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 293, 20 pages. doi:10.1145/3613904.3642111

[67] Jared Katzman, Angelina Wang, Morgan Scheuerman, Su Lin Blodgett, Kristen Laird, Hanna Wallach, and Solon Barocas. 2023. Taxonomizing and measuring representational harms: A look at image tagging. In *Proceedings of the AAAI Conference on artificial intelligence*, Vol. 37. 14277–14285.

[68] Sarah Kiden, Oriane Peter, Gisela Reyes-Cruz, Maira Klyshbekova, Sena Choi, Aislinn Gomez Bergin, Maria Waheed, Damian Eke, Tayyaba Azim, Sarvapali Ramchurn, et al. 2025. Back to the Communities: A Mixed-Methods and Community-Driven Evaluation of Cultural Sensitivity in Text-to-Image Models. *arXiv preprint arXiv:2510.27361* (2025).

[69] Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. *arXiv preprint arXiv:2205.10646* (2022).

[70] Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

[71] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2023. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867* (2023).

[72] Xingyu Lan, Jiaxi An, Yisu Guo, Tong Chiyou, Xintong Cai, and Jun Zhang. 2025. Imagining the Far East: Exploring Perceived Biases in AI-Generated Images of East Asian Women. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 332, 7 pages. doi:10.1145/3706599.3719678

[73] Seonghee Lee, Maho Kohga, Steve Landau, Sile O'Modhrain, and Hari Subramonyam. 2024. AltCanvas: A Tile-Based Editor for Visual Content Creation with Generative AI for Blind or Visually Impaired People. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 70, 22 pages. doi:10.1145/3663548.3675600

[74] Qisheng Li and Shaomei Wu. 2024. "I Want to Publicize My Stutter": Community-led Collection and Curation of Chinese Stuttered Speech Data. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 475 (Nov. 2024), 27 pages. doi:10.1145/3687014

[75] Zi-Ao Ma, Tian Lan, Rong-Cheng Tu, Shu-Hang Liu, Heyan Huang, Zhijing Wu, Chen Xu, and Xian-Ling Mao. 2025. T2I-Eval-R1: Reinforcement Learning-Driven Reasoning for Interpretable Text-to-Image Evaluation. *arXiv preprint arXiv:2505.17897* (2025).

[76] Kelly Mack, Rai Ching Ling Hsu, Andrés Monroy-Hernández, Brian A. Smith, and Fannie Liu. 2023. Towards Inclusive Avatars: Disability Representation in Avatar Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 607, 13 pages. doi:10.1145/3544548.3581481

[77] Kelly Avery Mack, Rida Qadri, Remi Denton, Shaun K. Kane, and Cynthia L. Bennett. 2024. "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* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 288, 23 pages. doi:10.1145/3613904.3642166

[78] Jabez Magomere, Shu Ishida, Tejumade Afonja, Aya Salama, Daniel Kochin, Yuehgoh Foutse, Imane Hamzaoui, Raesetje Sefala, Aisha Alaagib, Samantha Dalal, et al. 2025. The World Wide recipe: A community-centred framework for fine-grained data collection and regional bias operationalisation. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 246–282.

[79] Daniela Massiceti, Camilla Longden, Agnieszka Słowik, Samuel Wills, Martin Grayson, and Cecily Morrison. 2024. Explaining CLIP's performance disparities on data from blind/low vision users. arXiv:2311.17315 [cs.CV] https://arxiv.org/abs/2311.17315

[80] Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. 2021. Orbit: A real-world few-shot dataset for teachable object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10818–10828.

[81] Nusrat Jahan Mim, Dipannita Nandi, Sadaf Sumyia Khan, Arundhuti Dey, and Syed Ishtiaque Ahmed. 2024. 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*. 1–18.

[82] OpenAI. 2025. GPT-Image-1. https://platform.openai.com/docs/guides/image-generation?image-generation-model=gpt-image-1. Accessed: 2025-09-03.

[83] OpenAI. 2025. GPT4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2025-09-03.

[84] Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020. Masakhane–machine translation for Africa. *arXiv preprint arXiv:2003.11529* (2020).

[85] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. 2023. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14277–14286.

[86] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[87] Nikita Pavlichenko and Dmitry Ustalov. 2023. Best prompts for text-to-image models and how to find them. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2067–2071.

[88] Erin Pritchard. 2024. *Let's Tackle Representations of Dwarfism in the Media.* https://www.hope.ac.uk/news/allnews/lets-tackle-representations-of-dwarfism-in-the-media.html Liverpool Hope University News.

[89] Rida Qadri, Mark Diaz, Ding Wang, and Michael Madaio. 2025. The case for" thick evaluations" of cultural representation in ai. *arXiv preprint arXiv:2503.19075* (2025).

[90] Rida Qadri, Michael Madaio, and Mary L. Gray. 2025. Confusing the Map for the Territory. *Commun. ACM* (Sept. 2025), 4 pages. doi:10.1145/3731655 Online First.

[91] Rida Qadri, Renee Shelby, Cynthia L. Bennett, and Remi Denton. 2023. 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* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 506–517. doi:10.1145/3593013.3594016

[92] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[93] Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2023. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems* 36 (2023), 66127–66137.

[94] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*. Pmlr, 8821–8831.

[95] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.

[96] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[97] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems* 35 (2022), 25278–25294.

[98] JooYoung Seo, Sanchita S. Kamath, Aziz Zeidieh, Saairam Venkatesh, and Sean McCurry. 2024. MAIDR Meets AI: Exploring Multimodal LLM-Based Data Visualization Interpretation by and with Blind and Low-Vision Users. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 57, 31 pages. doi:10.1145/3663548.3675660

[99] Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. Dosa: A dataset of social artifacts from different indian geographical subcultures. *arXiv preprint arXiv:2403.14651* (2024).

[100] Summer S Shelton and T Franklin Waddell. 2021. Does 'inspiration porn'inspire? How disability and challenge impact attitudinal evaluations of advertising. *Journal of Current Issues & Research in Advertising* 42, 3 (2021), 258–276.

[101] Ben Shneiderman. 2021. Human-centered AI: A new synthesis. In *IFIP Conference on Human-Computer Interaction*. Springer, 3–8.

[102] Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619* (2024).

[103] Jonatan Södergren and Niklas Vallström. 2023. Disability in influencer marketing: a complex model of disability representation. *Journal of Marketing Management* 39, 11-12 (2023), 1012–1042.

[104] Hariharan Subramonyam, Colleen Seifert, and MI Eytan Adar. 2021. How can human-centered design shape data-centric AI. In *Proceedings of NeurIPS Data-Centric AI Workshop*. Curran Associates New York, NY, USA, 3.

[105] David Welch Suggs and Jason Lee Guthrie. 2017. Disabling Prejudice: A Case Study of Images of Paralympic Athletes and Attitudes Toward People with Disabilities. *International Journal of Sport Communication* 10, 2 (2017), 258–276. https://www.researchgate.net/publication/317286559_Disabling_Prejudice_A_Case_Study_of_Images_of_Paralympic_Athletes_and_Attitudes_Toward_People_With_Disabilities

[106] Cella M Sum, Franchesca Spektor, Rahaf Alharbi, Leya Breanna Baltaxe-Admony, Erika Devine, Hazel Anneke Dixon, Jared Duval, Tessa Eagle, Frank Elavsky, Kim Fernandes, et al. 2024. Challenging ableism: A critical turn toward disability justice in hci. *XRDS: Crossroads, The ACM Magazine for Students* 30, 4 (2024), 50–55.

[107] Mei Tan, Hansol Lee, Dakuo Wang, and Hari Subramonyam. 2024. Is a seat at the table enough? Engaging teachers and students in dataset specification for ml in education. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–32.

[108] Mei Tan, Hansol Lee, Dakuo Wang, and Hari Subramonyam. 2024. Is a Seat at the Table Enough? Engaging Teachers and Students in Dataset Specification for ML in Education. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 81 (April 2024), 32 pages. doi:10.1145/3637358

[109] Xinru Tang, Ali Abdolrahmani, Darren Gergle, and Anne Marie Piper. 2025. Everyday Uncertainty: How Blind People Use GenAI Tools for Information Access. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 63, 17 pages. doi:10.1145/3706598.3713433

[110] Anja Thieme, Aditya Nori, Marzyeh Ghassemi, Rishi Bommasani, Tariq Osman Andersen, and Ewa Luger. 2023. Foundation models in healthcare: Opportunities, risks & strategies forward. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–4.

[111] Rong-Cheng Tu, Zi-Ao Ma, Tian Lan, Yuehao Zhao, Heyan Huang, and Xian-Ling Mao. 2024. Automatic evaluation for text-to-image generation: Task-decomposed framework, distilled training, and meta-evaluation benchmark. *arXiv preprint arXiv:2411.15488* (2024).

[112] Jonathan Van Geuns, Ana Brandusescu, and Mozilla Insights. 2020. What Does it Mean? Shifting Power Through Data Governance. *Mozilla Foundation 2020)< https://foundation. mozilla. org/en/data-futureslab/data-for-empowerment/shifting-power-through-data-governance/> accessed 8 September 2025* (2020).

[113] Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2022. A Study of Implicit Bias in Pretrained Language Models against People with Disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 1324–1332. https://aclanthology.org/2022.coling-1.113/

[114] Susan Vertoont, Tina Goethals, Frederik Dhaenens, Patrick Schelfhout, Tess Van Deynse, Gabria Vermeir, and Maud Ysebaert. 2022. Un/recognisable and dis/empowering images of disability: A collective textual analysis of media representations of intellectual disabilities. *Critical Studies in Media Communication* 39, 1 (2022), 1–14.

[115] Hanna Wallach, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow, et al. 2025. Position: Evaluating Generative AI Systems is a Social Science Measurement Challenge. *arXiv preprint arXiv:2502.00561* (2025).

[116] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022. Measuring representational harms in image captioning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 324–335.

[117] Alison Wilde. 2022. The representation of disabled women and recent disabled women-led media. *Disability & Society* 37, 3 (2022), 522–527.

[118] Kexin Zhang, Elmira Deldari, Zhicong Lu, Yaxing Yao, and Yuhang Zhao. 2022. "It's Just Part of Me:" Understanding Avatar Diversity and Self-presentation of People with Disabilities in Social Virtual Reality. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 4, 16 pages. doi:10.1145/3517428.3544829

[119] Kexin Zhang, Edward Glenn Scott Spencer, Abijith Manikandan, Andric Li, Ang Li, Yaxing Yao, and Yuhang Zhao. 2025. Inclusive Avatar Guidelines for People with Disabilities: Supporting Disability Representation in Social Virtual Reality. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 560, 26 pages. doi:10.1145/3706598.3714230

[120] L. Zhang and B. Haller. 2013. Consuming Image: How Mass Media Impact the Identity of People with Disabilities. *Communication Quarterly* 61, 3 (2013), 319–334. doi:10.1080/01463373.2013.776988

[121] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3836–3847.

[122] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems* 36 (2023), 55006–55021.

# A Appendix

## A.1 Additional Interface Screenshots

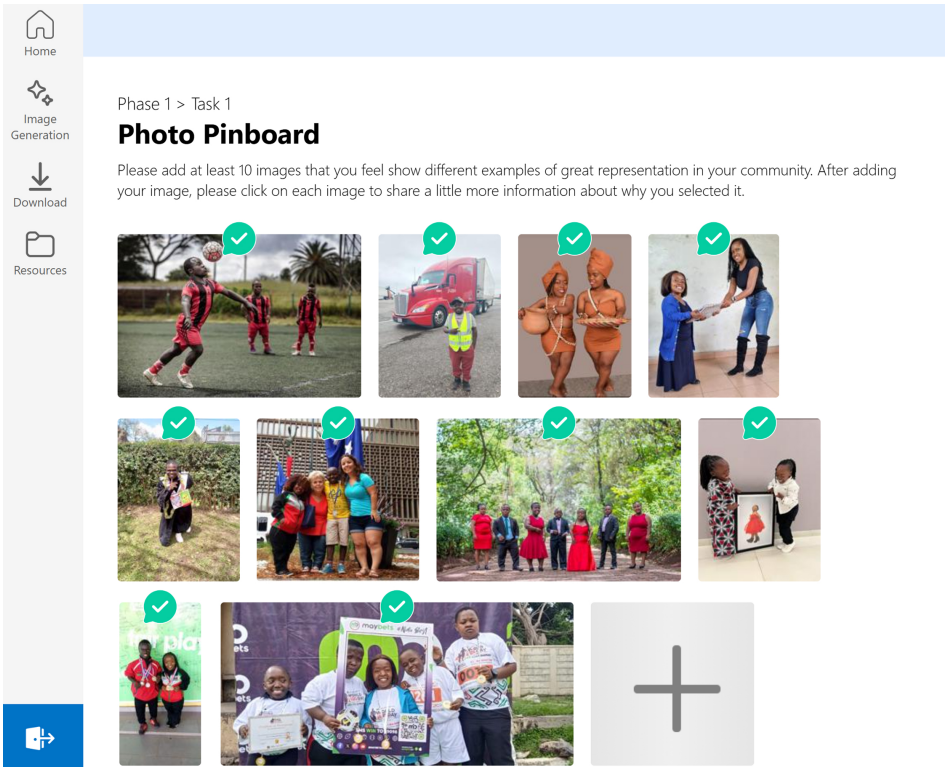


**Figure 6: Example of Pinboard activity landing page that asks project lead to select at least 10 images as different examples of great representation of their community.**

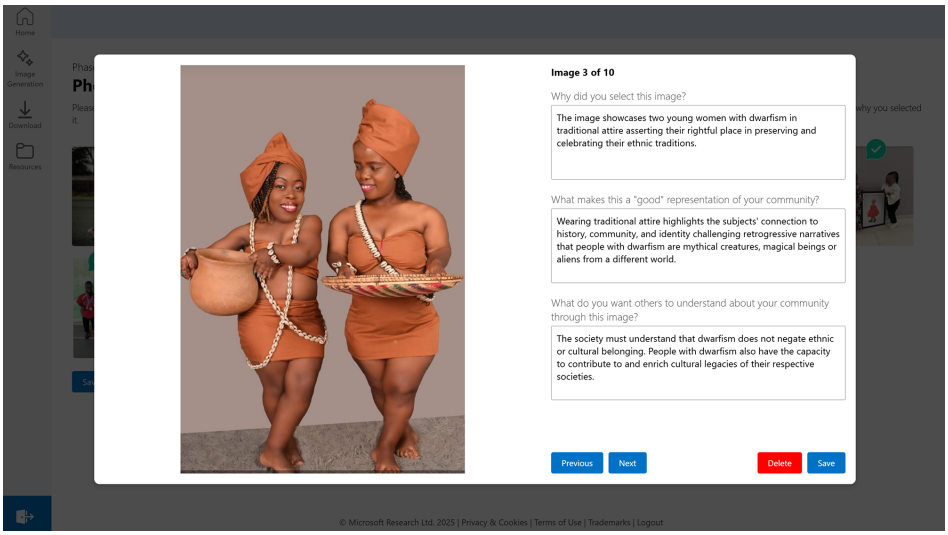


**Figure 7: This screenshot shows the three reflective questions that project leads were asked to complete for each image of the Pinboard.**

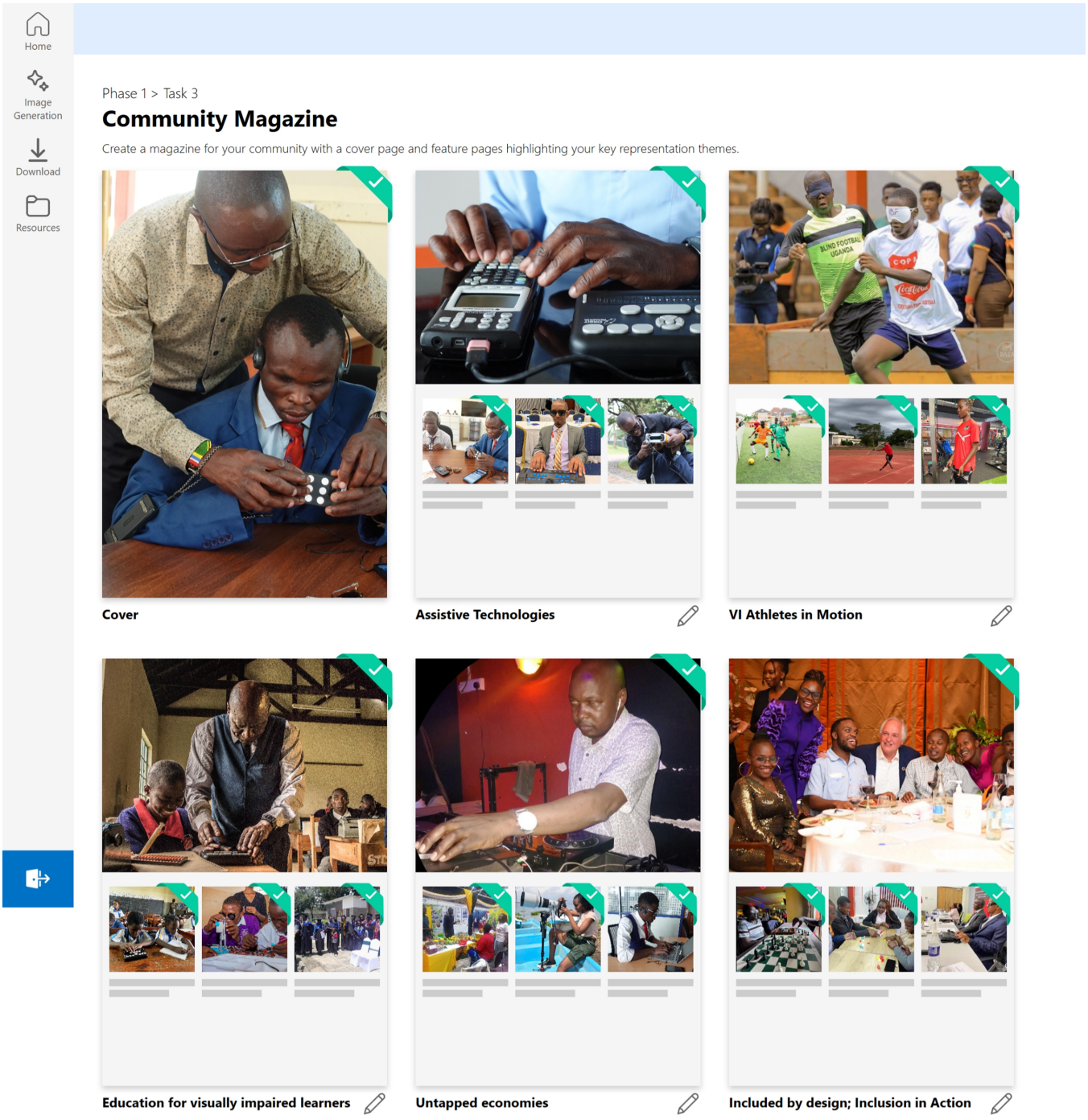


**Figure 8: Example of the Magazine activity completed by one of the organizations, showing the cover page image, as well as the five feature page titles and corresponding image selections to show main representational themes and their variations.**

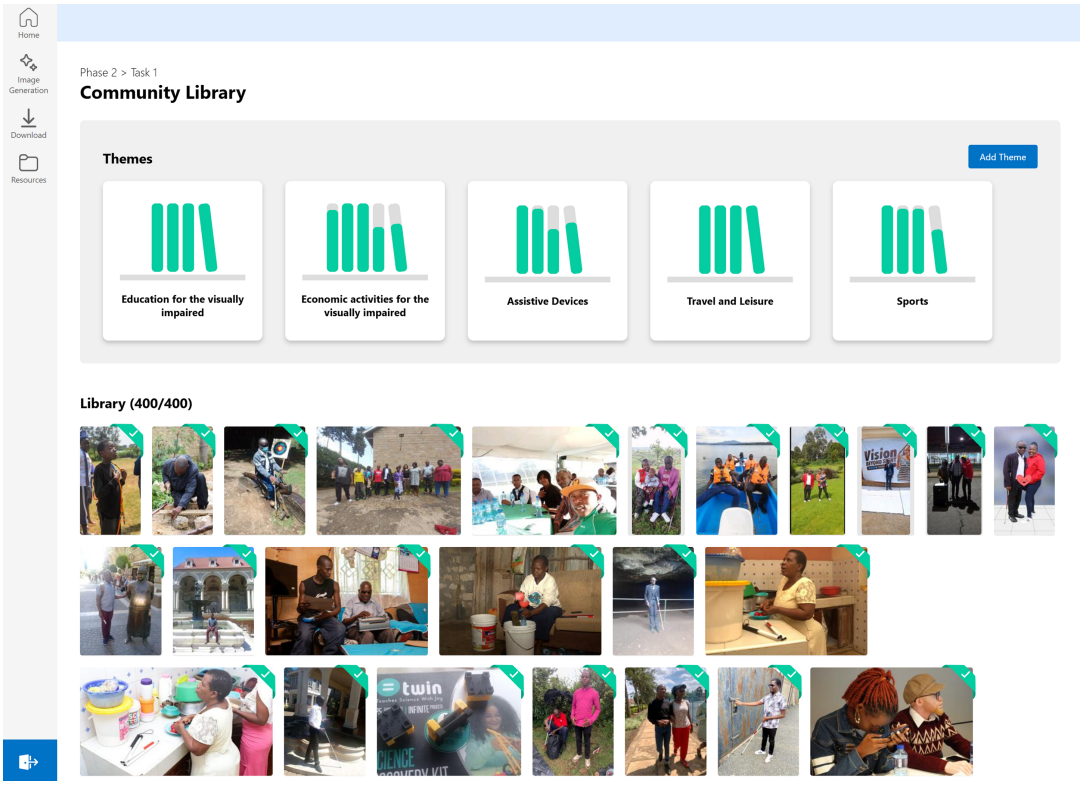


**Figure 9: This screenshot illustrates the high-level organization of a communities' thematic representation definition via the Community Library.**

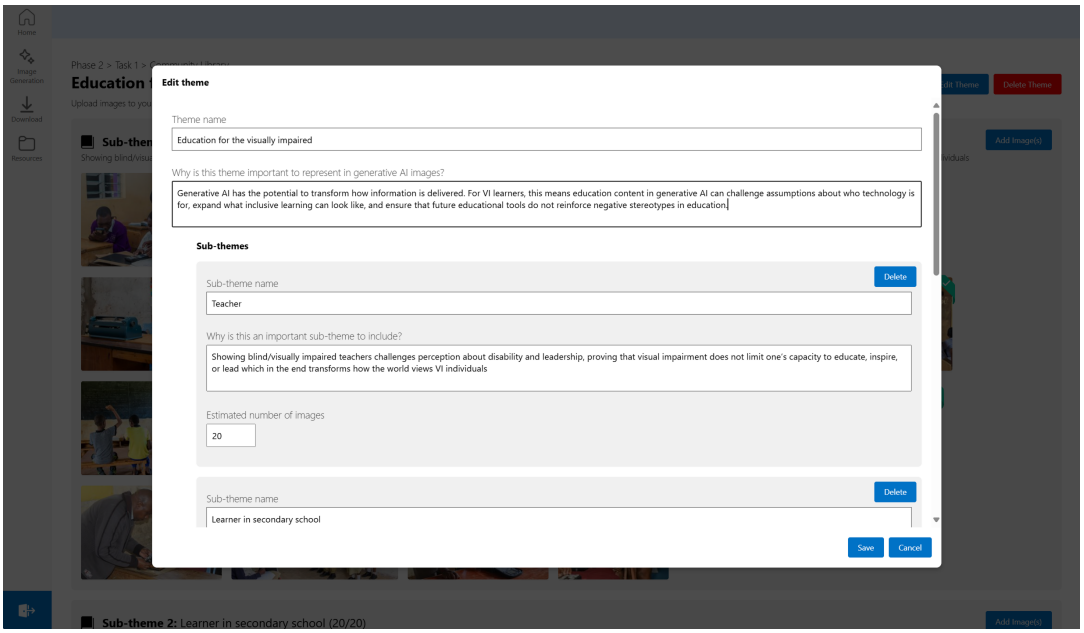


**Figure 10: This screenshot illustrates how project leads could edit, add, remove sub-themes,described the importance of including a specific sub-theme within the main theme and the number of images they estimated they need to achieve balance across their approx. 400 image dataset.**

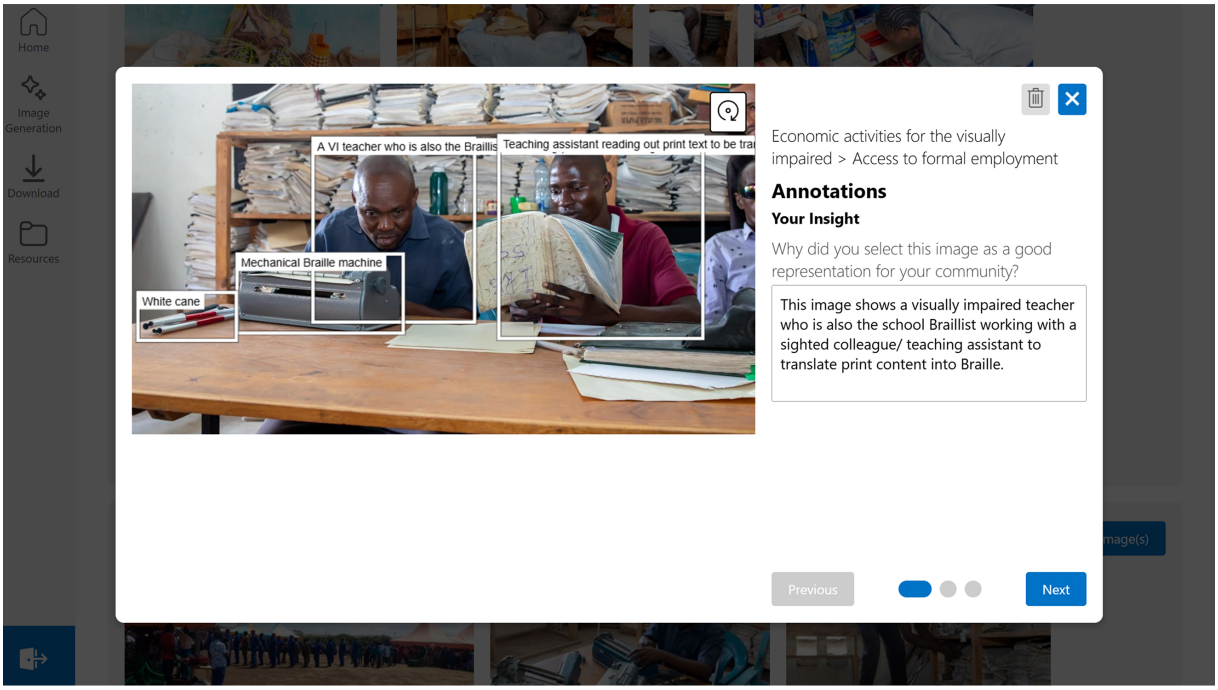


**Figure 11: This screenshot shows the first frame (1/3) of the per-image data annotation sequence: project leads are asked to explain why the selected the image as a good representation for their community.**

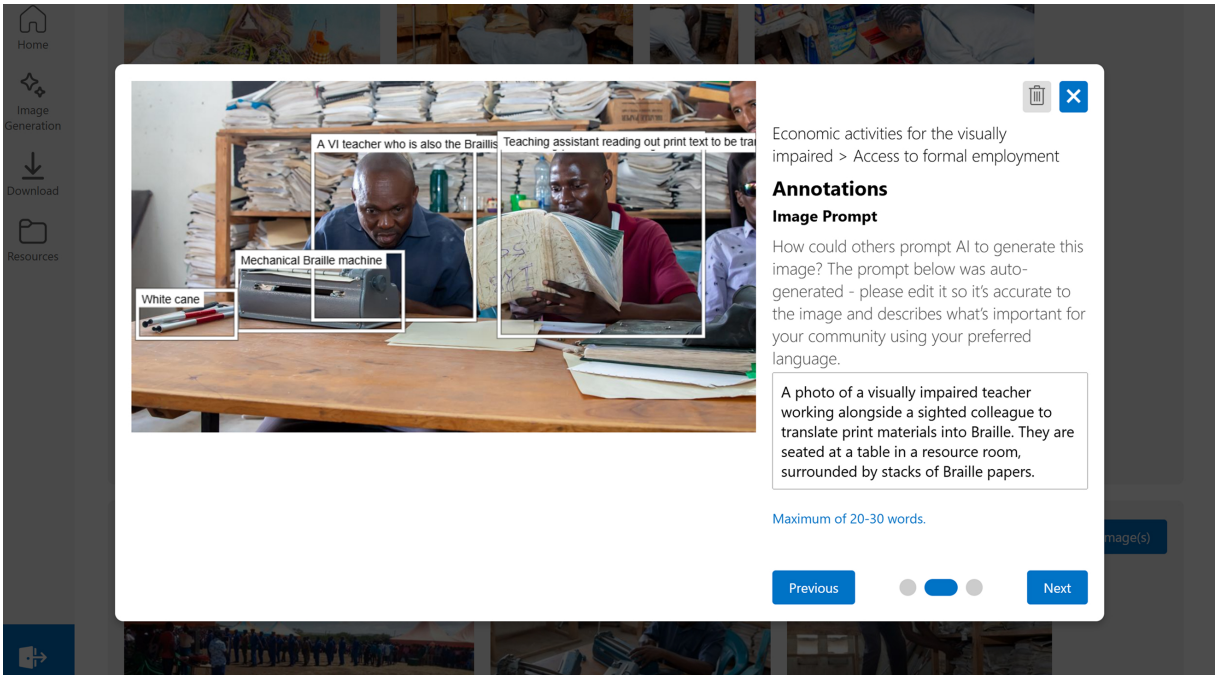


**Figure 12: This screenshot shows the second frame (2/3) of the per-image data annotation sequence: project leads review and can make any necessary edits to the displayed prompt that is auto-generated using the image and their initial explanation for selecting it as inputs to the GPT4o model.**

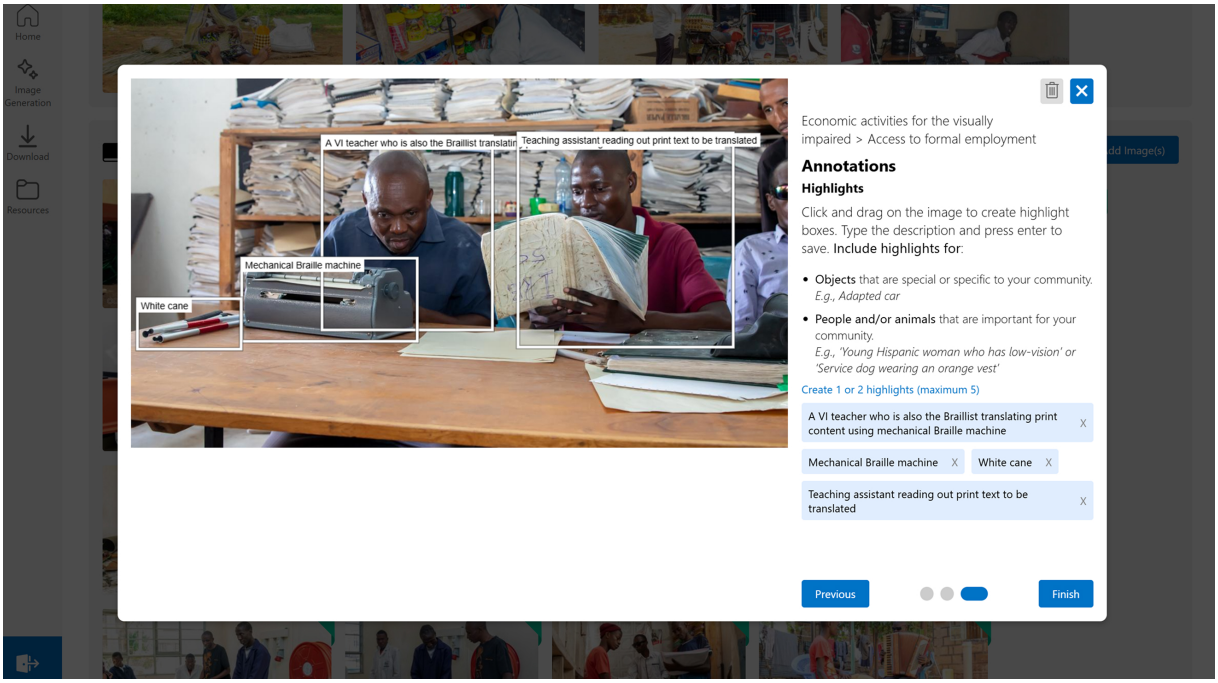


**Figure 13: This screenshot shows the third frame (3/3) of the per-image data annotation sequence: it details the instructions given to project leads for how to highlight key 'objects' and 'people/ animals' of importance to their community via labeled bounding boxes.**

## A.2 Example of a Thematic Community Library Structure

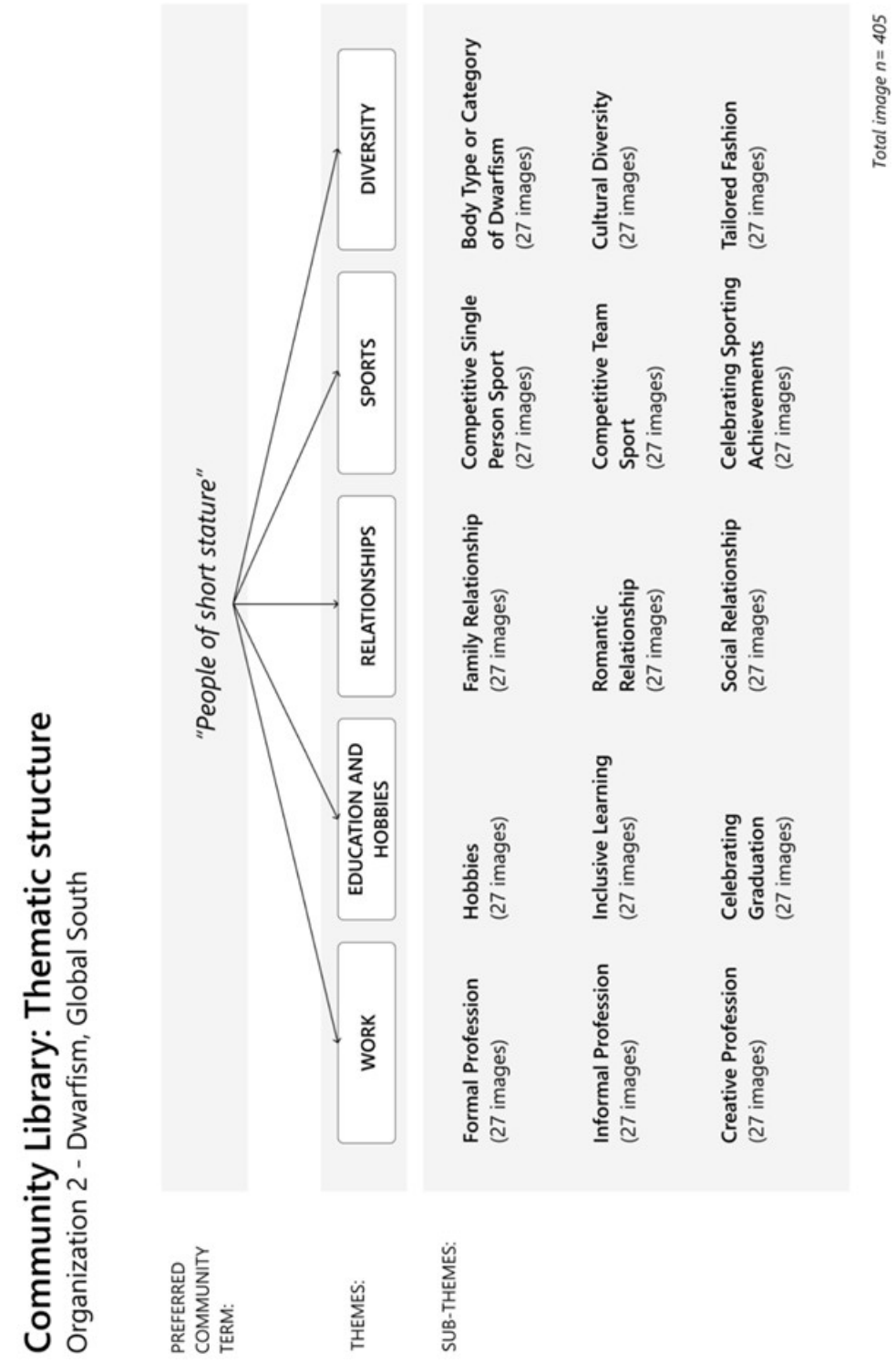


**Figure 14: Schematic illustration of the thematic Community Library structure of one of the disability organizations. It shows: (i) the communities preferred community term, (ii) five main themes; and (iii) their respective sub-themes – including numerical indications of how many images were included in each representation category.**

## A.3 Additional AI Evaluation Examples

In this section, we include two more qualitative examples of how project leads were evaluating AI generated images as part of their rating task.

The image in Figure 15 was rated as "very good" by the project lead representing the vision impairment community (P1), who considered several factors: the *realism* of the image; how well it *aligned with the prompt* (e.g., if it depicted a computer class); the overall *authenticity* of the composition, and *accuracy of the AT* (e.g., looks like the Orion apart from sound-enabled beats); as well as asssessments of how clearly the *visual impairment is recognizable* and *appropriately* shown. The commentary also suggests the image could be further improved by more clearly illustrating a mixed-ability classroom environment, moving it closer to an ideal representation. In P1's own words:

> "And this is a very realistic sort of picture. The only thing that I'll wonder is, and I know in some of our datasets we have a calculator there, but if it's a computer class, sometimes we usually do not need a calculator, but that calculator looks like the Orion [accessible calculator] apart from having the sound enabled beats. But the image itself looks really nice in the composition and the guy in the center, you can just by looking at that person, you can see they have some sort of visual impairment which makes a lot of sense. And also the learners who are using the earphones that makes sense. So for contrast, or maybe an add-on, is maybe having someone who are low vision, so just to complement the composition so that somebody can tell there's a difference between which learner is wearing earphones and the other one is not wearing earphones. These can just be some of the subtle kind of tell sign, but generally the image really looks well composed." (P1, wk12, meeting)

The AI-generated image in Figure 16 was rated as "good" by the project lead of the Global South dwarfism community (P2). Asking what held them back from giving it a "very good" score, P2 highlighted that the image *appropriately* depicted the *physical body proportions* and commended the overall *realism* of the image by describing it as "not hyper realistic" – which P2 explains to us as reflecting a positive statement that refers to the good, *natural* depiction of the image with "no additives or sugar coating". However, the choice of attire — a t-shirt instead of a clear sporting outfit — made the scene appear less *believable* for a competition setting with a trophy and medal. In P2's own words:

> "Excellent interpretation of the physical attributes of a person of short stature. But the image is not hyper realistic" (P2, wk12, system entry)
>
> "[Making reference to another sporting image that was assessed as "very good"] The difference in proportion is very small. That's what I can say, frankly, because of what the person is wearing. It shows like the person is wearing just a casual t-shirt. Not having probably a name or a flag or the t-shirt is very, it doesn't show like it's a sporty t-shirt, but it shows the attire." (P2, wk13, meeting)

"A photo of a group of visually impaired learners in secondary school working in a computer lab. They are using laptops equipped with screenreader software for their studies."

**Figure 15: Example of an image-prompt pair rated as "very good" in the AI evaluation task.**

"A photo of a person of short stature proudly holding a winner's trophy and wearing a medal around this neck on a sports court."

**Figure 16: Example of an image-prompt pair rated as "good" in the AI evaluation task.**