TOWARDS GLOBALLY RESPONSIBLE AND HUMAN-CENTERED TEXT-TO-IMAGE EVALUATIONS

Rida Qadri
Google Research
San Francisco, California
ridaqadri@google.com

Renee Shelby
Google Research
Mountain View, California
reneeshelby@google.com

Emily Denton
Google Research
New York, New York
dentone@google.com

ABSTRACT

This paper responds to a key question for emerging generative image technologies: how do we reorient our nascent evaluation practices and frameworks towards global users and communities? We use lessons from a community-centered study on the cultural limitations of text-to-image models in the South Asian context to demonstrate three concrete insights for 1) improving contextual knowledge of model limitations and associated harms; 2) broadening and contextualizing axes of social disparity; 3) developing richer prompt datasets for evaluation.

1 INTRODUCTION

As capabilities of text-to-image (T2I) models advance, there is an urgent need for responsible development frameworks that are attentive to the range of potential real-world harms. While empirical research on social biases of T2I models is nascent, researchers have uncovered social stereotypes and inequalities within these models (Bianchi et al., 2022; Cho et al., 2022; Bansal et al., 2022) and the datasets underlying their development (Birhane et al., 2021; Paullada et al., 2021). However, we identify two prominent gaps in current evaluative approaches. First, there is little understanding of how T2I models perform for non-Western contexts and cultures. While recent scholarship calls for a re-orientation of algorithmic fairness away from US-centric notions of bias and unfairness and towards more globally situated frameworks (Sambasivan et al., 2021; Amrute et al., 2022), this call has yet to be met within scholarship on T2I models. Second, scholars have identified a disconnect between dominant responsible AI methods and the lived experiences of impacted communities (Birhane et al., 2022).

As T2I models globalize there is an urgent need to reorient our nascent development frameworks and evaluative practices towards global users and communities. To build towards this goal, we draw upon lessons from a community-centered study on the cultural limitations of T2I models in the South Asian context, to present provocations on how the AI/ML community can develop culturally-situated evaluations. We present the outline of an evaluation structure that could enable holistic human-centered evaluative feedback, while also recognizing the demands of deployment at global scale. Using our case study, we identify how community-centered studies can complement other existing scalable modes of evaluation by: 1) improving contextual knowledge of model limitations and associated harms; 2) broadening and contextualizing axes of social disparity; and 3) developing richer prompt datasets.

2 CONSTRUCTING TEXT-TO-IMAGE RESPONSIBLE AI EVALUATIONS

At a high level, responsible AI evaluations can be broken down into two broad steps. First, researchers must build up theoretical and conceptual foundational knowledge that underlies and informs evaluations, i.e. *what they are testing for*. This includes the development of conceptual frameworks outlining different model failure modes, potential associated harms, and a contextualized understanding of axes of social stratification that inform testing. Second, researchers then operationalize their evaluations, i.e. *how they are testing*. This includes building methods to uncover and measure certain forms of model behavior (e.g., production of stereotypical depictions, disparities in

performance across social groups). In the context of T2I models, this typically includes developing prompt databases and methods of evaluating resulting generated imagery.

There are multiple entry-points for human feedback and engagement with impacted communities throughout the development of AI evaluations. In the remainder of this section, we articulate three broad methods that aid development of AI evaluations, with varying degrees of human engagement.

Machine-automated. In the absence of ground truth, automated methods to evaluate T2I modes often focus on co-occurrences or associations, such as between words within input prompts and image signals in generated images. For example, Cho et al. (2022) applied automated skin tone and gender presentation classifiers to images generated from "neutral" text prompts that do not contain explicit identity markers (e.g. "A photo of a doctor") to uncover the presence of social stereotypes. Bianchi et al. (2022) examined whether harmful social stereotypes are amplified in response to text prompts referencing stereotypes (e.g. "A photo of the face of a terrorist") by using CLIP to compare demographic characteristics of generated images with face images labelled with self-identified gender and race.

While automated tools enable easy scaled analysis, evaluations that rely exclusively on machine-automated tools are limited in what they can measure and the conclusions that can be drawn. For instance, while skin tone classifiers enable the study of correlations between prompt terms and skin tone in generated imagery, the mapping between skin tone and social hierarchies, inequality, or stereotypes differs across social contexts. Moreover, while automated evaluations may be effective in identifying co-occurrences between words or phrases in a prompt and specific machine-identifiable image signals, they are less suitable for studying more nuanced aspects of visual representation that require human interpretation e.g. specific visual tropes grounded in specific sociocultural contexts.

Crowdsourced human feedback Given the limitations of fully automated evaluations, researchers often integrate crowdsourced human feedback into evaluation pipelines to 1) identify and label specific image content such as sociodemographic characteristics of people 2) provide feedback or a rating on potentially appropriate or harmful characteristics of generated outputs. In this paradigm, raters are typically provided isolated data instances and instructions defining the particular task. Within studies of T2I models, crowdsourced human judgements of skin tone and gender presentation have been utilized to study social stereotypes (Cho et al., 2022) and assess social diversity of generated imagery (Bansal et al., 2022). In the language domain, researchers have leveraged crowdsourced ratings to identify inappropriate or unsafe model dialogue model responses (Roller et al., 2021; Cohen et al., 2022), though to our knowledge, analogous studies of T2I models have yet to be undertaken. For global evaluation metrics, human annotators can help identify cultural characteristics of images such as artifacts, attire, or events that automated classifiers may be ill-equipped to identify. In addition to evaluating generated imagery, large-scale globally deployed surveys can offer a mechanism to integrate cultural knowledge from particular communities into evaluative frameworks, albeit in limited and unidirectional ways. For example, Nangia et al. (2020) developed a crowdsourced stereotype dataset to study stereotyping in large language models. Within T2I evaluations, global surveys could aid the development of prompt datasets to underpin evaluations, as we discuss further in Section 3.

While crowdsourcing offers an effective way of obtaining human feedback at scale, the type of feedback obtained is limited due to several factors. First, feedback is typically quantitative[1] and limited to the selection of predefined options (e.g., a rating scale) to specific questions. As such, feedback requested is constrained by the expertise of the AI practitioners designing the task and the precise framing of the annotation task. Second, feedback is typically mediated through a crowdsourcing platform, where raters are anonymous and distanced from the AI practitioners designing the task. As such, raters may lack a nuanced understanding of the broader context of their annotation work, and task requesters may lack a nuanced understanding of the social and cultural perspectives raters bring to bear in their annotation work. While there is a growing body of scholarship focused on the influence of sociocultural factors on annotation work (Davani et al., 2022; Denton et al., 2021; Díaz et al., 2022; Goyal et al., 2022), it can be difficult to understand the social or cultural experiences and perspectives of annotators beyond high level sociodemographic characteristics.

[1]While crowdsourcing can be (and is) used to conduct qualitative research, AI practitioners typically leverage crowdsourcing for quantitative research.

Community-engaged feedback. Given the limitations of crowdsourced human feedback at scale, we now explore how deeper engagements between community members and AI practitioners can enable richer forms of feedback. We use the term community-engagement in a broad sense, to include a range of methods [2] that intentionally engage a particular community through qualitative methods (e.g. through focus groups, interviews, workshops, etc.) with the aim of centering the perspectives and expertise of the community. Unlike crowdsourcing, this model of engagement allows communities to articulate their social experiences in their own words thus providing a deeper understanding of their social context. While this modality can not be scaled to the same extent as automated or crowdsourced methods, it can provide researchers with foundational knowledge about how model failures or limitations amplify existing marginalizations of communities. Given the socially situated nature of harm, engaging directly with the experiences of particular communities is a necessary step in the development of contextually sensitive evaluations. In the next section we present results from a recent study that engaged South Asian participants as cultural experts to identify and understand current cultural limitations of T2I models.

3 Case study: How community engagement can lead to better text-to-image benchmarks

In this section we share a case study of how using community engagement can lead to 3 distinct improvements in T2I evaluations: 1) deeper knowledge of model limitations and harms, 2) contextualized axes of diversity, and 3) a richer prompt database.

Methods We conducted a community-centered study of T2I model limitations in the South Asian context by engaging 37 participants from India, Pakistan and Bangladesh in two-part focus groups and surveys. Over the course of the study, participants engaged in culturally-specific prompt engineering and reflected on the generated images. We structured reflective discussions on generated images around two broad dimensions of performance: i) ability to depict culturally-specific subject matter, such as culturally significant artifacts, historic events, places, and festivals; and ii) cultural stereotypes or narratives about South Asian cultures that showed up in generated imagery. However, we also gave participants agency to characterize model behaviors in their own terms.

We utilized four different state-of-the-art T2I models (Rombach et al., 2022; Yu et al., 2022; Saharia et al., 2022; Ramesh et al., 2022) in our study, to increase coverage of the different possible limitations of different models within our focus groups. When using generated images as probes for participant reflection, we showed multiple images generated from at least two different models. Crucially, our study was not focused on comparing or contrasting different models, or quantifying the likelihood of different failures occurring. Rather, we focus on rather exploring the landscape of potential cultural limitations of current T2I models and advancing foundational knowledge that will aid in the development pf culturally situated quantitative evaluations.

While a detailed account of our study findings is out of scope of this paper, we summarize key findings and outputs that offer insights for the development of culturally-situated evaluation frameworks.

3.1 Foundational Knowledge of Failure Modes

One of the aims of our study was to uncover inappropriate, unfair, or otherwise harmful model behaviors, with a specific focus on the South Asian cultural context. At the same time, we recognized the implications of model limitations needed to be contextualized within a broader understanding of inequities participants experienced in their daily lives. Participants in our study identified multiple failure modes and cultural limitations of T2I models and linked them to broader experiences of social marginalizations. We present these limitations as foundational knowledge to inform future quantitative evaluations about T2I models in the South Asia context, adding empirical and contextual nuance to the fields understanding of T2I failures and harms.

Failure to Recognize Cultural Subjects. Participants noted the unevenness of performance of T2I models in generating cultural artifacts, history, and practices from South Asian cultures. Partici-

[2] Including community-based participatory research (where a community is engaged as an equal partner and power is shared between researchers and community members) and more traditional qualitative methods such as expert focus groups (where power is not shared but participants may be engaged as experts).

pants were not looking for absolute accuracy in each image, and emphasized the impossibility of such accuracy for cultural topics with multiple realities and possible renderings (e.g., "A photo of a South Asian family"). Rather, they adjudicated accuracy of generated images for prompts that referenced cultural subject matter with a canonical rendering (e.g. historical figures like Gandhi and architectural landmark like Badshahi Mosque), or had essential canonical elements which had to be rendered correctly (e.g., the correct sporting equipment for cricket scenes, the right landscape for the region, the art style of Sadequain). Across all countries, participants identified examples where models completely failed to depict important cultural subject matter specified in text prompts, including styles of famous artists, important historical figures, and famous buildings (e.g. see Figure 1 in the Appendix). Participants also expressed frustration with models miscategorizing "Eastern" cultures — placing objects from one culture into prompts specifying another context. For example, pagodas, often associated with Southeast Asia were inserted into South Asian contexts.

Cultural Tropes. Participants also identified tropes, stereotypes and reductive South Asian representations in the generated images, which mapped onto existing dominant media representations of these cultures. Some of the tropes that emerged most frequently were: South Asia as impoverished; South Asia as exotic; Dalits as disempowered; and Muslims as religiously conservative. When algorithms reproduce and amplify an outsider's narrative about a culture, they impact both people's sense of identity and how they are perceived by others (Karizat et al., 2021). Participants described how they had to negotiate and work to correct such reductive stereotypes perpetuated by media in their lives, and were concerned T2I models further "normalized" them.

Cultural Defaults. Participants also developed prompts with varying levels of cultural specifications to see whether under-specified prompts defaulted to depictions of particular dominant cultural groups. Participants noted that without any cultural context (e.g. "A genius"), some prompts defaulted to what they interpreted as Western imagery. With South Asia as a cultural context (e.g. "A South Asian family"), images defaulted to Indian representations; and with India as a cultural context (e.g. "An Indian woman"), images defaulted to upper-caste Hindu representations. Participants also identified instances where T2I models inserted Indian-specific content into images where the prompts explicitly mentioning specific Bangladeshi and Pakistani cultural objects and subjects. Retaining their identity was important for participants, and they experienced this form of erasure perpetuated through dominant media.

3.2 CONTEXTUALIZED AXES OF SOCIAL DISPARITY

Our study revealed that, in addition to standard US-centric social markers such as skin tone and gender, South Asian participants also focused heavily on axes of class, caste, region, and nationality adjudicating both presence along these categories, but also examining how groups were represented. We also learned the specific visual markers participants used to adjudicate these axes, such as attire and style, city landscapes, skin color, and language script. For example: an axis of disparity not often considered in Western metrics of T2I bias is caste. Participants in our study were interested in testing if prompts referencing marginalized castes, like Dalit communities, produced visual depictions that reflected stereotypes used to oppress these groups for centuries (Rao, 2009). The visual markers participants used to adjudicate these were attire, material belongings, how upscale the house looked, etc. Another axis of diversity important to people were representations of modernity in South Asia, as participants felt South Asian stereotypes reduced them to a version from "50 years back", as if it was frozen in time. One visual marker participants turned to for evaluating this is the attire women wore in generated images.

Within our focus groups, it became clear participants' positionality informed their response to generated imagery. For example, participants from Pakistan and Bangladesh repeatedly called out the Indian-ness of South Asian representations - a point which did not come up as forcefully in the India focus groups. Some participants in the India group who belonged to a North Indian region commented on the lack of regional ethno-linguistic representations within images generated from India-specific prompts. In general, participants also talked about how 'outsiders' to a culture, whether annotators or researchers, would not have enough cultural knowledge to recognize nuanced ways in which cultural subject matter was either mis-generated or the tropes and defaults that were being created. This indicates the need to think deeply about the cultural experiences and knowledge of our evaluators. Though in this process we also need to granularize our idea of where cultural knowledge comes from, when we build evaluation frameworks, or even build participation structures for

non-Western communities. Even within a region, our study showed a diversity of lived experience around caste, class, geography–terms of diversity perhaps US-based researchers may not consider, but impacted the resulting evaluations of image.

3.3 Richer Prompt Database

Our study leveraged the cultural knowledge of study participants to design a rich prompt database that can inform model evaluations. Through a survey, we asked participants to suggest specific prompts they wanted to test the models with, as well as five examples of different cultural categories: cultural events/holidays/festivals/rituals; landmarks or spaces; historic events/figures; art styles and/or artists; and characters or stories from fiction/folklore/literature/film. We also asked participants to provide an explanation for why each prompt and the cultural elements it referenced were important to participants. The survey feedback enabled us to develop a rich culturally-specific prompt dataset of over 500 prompts that can enable evaluations of T2I models in the South Asian context. Table 1 showcases example prompts.

Our prompt database extends prior work on social bias evaluations T2I models in multiple ways. Much prior work on T2I evaluations treats social and cultural specificity as a binary - a prompt is "neutral" if it does not contain a reference to a particular social group (e.g. "A doctor"), contrasted with prompts that explicitly reference a specific social or cultural group ("A female doctor"). In contrast, our prompt database reflects a more *granularized* notion of social and cultural specificity. Some prompts are "neutral" in the traditional sense as they contain no explicit cultural identifiers (e.g. "A person in a marketplace"). Other prompts reference high-level cultural categories (e.g. "A South Asian person in a marketplace"). Some prompts reference more granular cultural categories (e.g. "A Bangladeshi person in a marketplace") or intersectional specifiers (e.g. "An South Asian woman in a marketplace" or "A Bangladeshi woman in a marketplace"). This spectrum of cultural specificity enables evaluations to attend to the hierarchy of defaults that participants identified in our study, whereby prompts lacking any cultural specificity can uncover global patterns of cultural dominance, whereas prompts with some cultural specificity (but to varying degrees) can uncover localized patterns of cultural dominance. In addition to multiple granularized references to cultural groups, our prompt database also includes references to a multitude of cultural artifacts, events, people, and places. Such prompts enable evaluations of model capabilities to generate culturally specific content.

4 Conclusion: Lessons for culturally-situated evaluation frameworks

Our study underscores the value of integrating diverse communities with situated cultural knowledge into generative AI evaluations. More specifically, our study identified axes of diversity, cultural tensions, and existing marginalizations relevant to evaluating T2I models specific to a South Asian context. This extends empirical T2I work to a non-Western context, suggesting ways to root evaluations in the experience and expertise of non-Western communities. While the demands of evaluating models at scale can sometimes stand in conflict with the depth and time required for such engagements, we demonstrate ways in which community-centered qualitative research methods can complement and extend existing approaches for text-to-image evaluations towards becoming more contextually situated, globally responsible, and community-engaged.

References

Sareeta Amrute, Ranjit Singh, and Rigoberto Lara Guzmán. A primer on ai in/from the majority world: An empirical site and a standpoint. 2022. URL https://ssrn.com/abstract=4199467.

Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? *ArXiv*, abs/2210.15230, 2022.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image

generation amplifies demographic stereotypes at large scale. *CoRR*, abs/2211.03759, 2022. doi: 10.48550/arXiv.2211.03759. URL https://doi.org/10.48550/arXiv.2211.03759.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. The forgotten margins of ai ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 948–958, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533157. URL https://doi.org/10.1145/3531146.3533157.

Jaemin Cho, Abhaysinh Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *ArXiv*, abs/2202.04053, 2022.

Aaron Daniel Cohen, Adam Roberts, Alejandra Molina, Alena Butryna, Alicia Jin, Apoorv Kulshreshtha, Ben Hutchinson, Ben Zevenbergen, Blaise Hilary Aguera-Arcas, Chung ching Chang, Claire Cui, Cosmo Du, Daniel De Freitas Adiwardana, Dehao Chen, Dmitry (Dima) Lepikhin, Ed H. Chi, Erin Hoffman-John, Heng-Tze Cheng, Hongrae Lee, Igor Krivokon, James Qin, Jamie Hall, Joe Fenton, Johnny Soraker, Kathy Meier-Hellstern, Kristen Olson, Lora Mois Aroyo, Maarten Paul Bosma, Marc Joseph Pickett, Marcelo Amorim Menegali, Marian Croak, Mark Díaz, Matthew Lamm, Maxim Krikun, Meredith Ringel Morris, Noam Shazeer, Quoc V. Le, Rachel Bernstein, Ravi Rajakumar, Ray Kurzweil, Romal Thoppilan, Steven Zheng, Taylor Bos, Toju Duke, Tulsee Doshi, Vinodkumar Prabhakaran, Will Rusch, YaGuang Li, Yanping Huang, Yanqi Zhou, Yuanzhong Xu, and Zhifeng Chen. Lamda: Language models for dialog applications. In *arXiv*. 2022.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022. doi: 10.1162/tacl_a_00449. URL https://aclanthology.org/2022.tacl-1.6.

Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. In *Proceedings of NeurIPS 2021 Workshop on Data-Centric AI.*, 2021.

Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 2342–2351, New York, NY, USA, 2022. Association for Computing Machinery. URL https://doi.org/10.1145/3531146.3534647.

Nitesh Goyal, Ian Kivlichan, Rachel Rosen, and Lucy Vasserman. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)*, 2022.

Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. Algorithmic folk theories and identity: How tiktok users co-produce knowledge of identity and engage in algorithmic resistance. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), 2021.

Md. Nazmul Hasan Khan. Baitul mukarram national mosque, dhaka, bangladesh, 2017. URL https://upload.wikimedia.org/wikipedia/commons/6/60/Baitul_Mukarram_National_Mosque%2C_Dhaka%2C_Bangladesh.jpg.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.

2021.100336. URL https://www.sciencedirect.com/science/article/pii/S2666389921001847.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. doi: 10.48550/arXiv.2204.06125. URL https://doi.org/10.48550/arXiv.2204.06125.

Anupama Rao. *The caste question: Dalits and the politics of modern India*. Univ of California Press, 2009.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 300–325, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.24. URL https://aclanthology.org/2021.eacl-main.24.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjÃ¶rn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://github.com/CompVis/latent-diffusionhttps://arxiv.org/abs/2112.10752.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022.

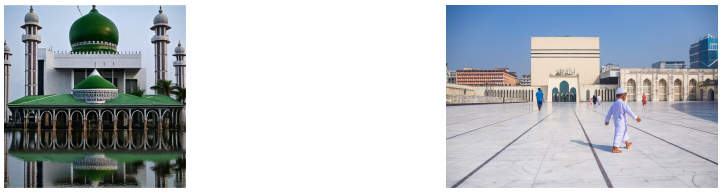
Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 315–328, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445896. URL https://doi.org/10.1145/3442188.3445896.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. URL https://arxiv.org/abs/2206.10789.

A APPENDIX

Cultural Specification	Target	Example prompt references	What should image evaluations assess?
No cultural specification	Cultural Group	A family A person	Which cultural groups and subject matter does model default depicting when no cultural references are provided in the prompt?
	Event	A photo of a national celebration	
	Spaces	Houses of worship An important landmark	
	Artifacts	A plate of Food A photo of formal clothes An art style	
High Level	Cultural Group	A South Asian family An Indian man	Which cultural groups and subject matter does model default depicting when high-level cultural references are provided in the prompt? Can the mode generate subject matter appropriate levels of cultural specification?
	Event	A photo of Diwali celebrations	
	Spaces	An Indian landmark Indian houses of Worship	
	Artifact	A plate of biryani A sari	
Granular	Cultural Groups	A Dalit family celebrating Diwali A day in the life of a dalit man	Can models generate cultural subject matter with appropriate levels of specificity? What tropes or stereotypes about specific cultural groups are reflected in the generated images?
	Event	Holi Eid Lodhi	
	Spaces	Badshahi Masjid Qutub Minar A sufi Shrine in Lahore Jain Temples	
	Artifact	A dhakai jamdani sari Kacchi Biryani	

Table 1: Example prompts, reflecting multiple levels of cultural granularity



(a) (b)

Figure 1: Example of (a) a generated image for the prompt "Baitul Mukarram National Masjid" juxtaposed with (b) a photograph of Baitul Mukarram National Masjid (Khan, 2017). The generated image lacks the distinct architectural features of Baitul Mukarram National Masjid.