Breaking the Binary: Addressing Gender Data Gaps in AI Systems

Pallab Mandal[1,2], Swaroop Srisailam[2], Praveen Pankajakshan[3], and Kumar Rajamani[2]
[1]Department of Mathematics and Scientific Computing, Indian Institute of Technology Kanpur, India
[2]CropIn Technologies, Bangalore, India
[3]Amrita Vishwa Vidyapeetham, Kerala, India
Emails: pallabm22@iitk.ac.in, swaroop.srisailam@cropin.com,
praveenpankaj@ieee.org, kumar.rajamani@cropin.com

***Abstract*—Individuals who identify themselves as non-binary, transgender and gender-nonconfirming are critically underrepresented in existing data sets, which hinders the development of equitable AI models. An analysis of data sets such as Conceptual Captions, Flickr30k, Multi30k, and COCO reveals minimal or no representation, with the result that the state-of-the-art architectures like Vision Transformers (ViT), ResNet, YOLO, and U-Net fall short in classifying, detecting, or segmenting men, women, and transgender individuals. While addressing this data gap is essential for creation of personalized AI solutions that are inclusive and equitable for all, it is also a delicate and multifaceted challenge. Bridging this gap involves ethical complexities, including the risk of reinforcing stereotypes through annotation practices that lack inclusivity. Additionally, the release of datasets without explicit consent or appropriate safeguards raises concerns regarding privacy, misuse, and potential harm to marginalized communities. This article seeks to initiate a critical dialogue on these issues rather than propose a definitive solution. We emphasize the importance of approaching this challenge with care, prioritizing ethical considerations and inclusivity. By fostering collaborative discussions among researchers, policymakers, and practitioners, we aim to explore pathways for developing AI systems that respect and reflect the full spectrum of human diversity without compromising privacy or safety.**

***Index Terms*—Transgender representation, computer vision, data set creation, ethical AI, inclusion, active learning, human-in-the-loop, open-source**

I. Introduction

The increasing reliance on Artificial Intelligence (AI) in critical applications such as healthcare, public safety, and autonomous systems assumes that the models will work equitably for all demographic groups. However, there is a glaring disparity persisting in the representation of individuals, who externally confirm as transgender individuals, within existing computer vision data sets. This under-representation of diverse gender identities not only reinforces biases, but also hinders the reliable performance and deployment of fair AI solutions in real-world scenarios where inclusion is of paramount importance.

The marginalized individuals often face systemic discrimination in societal structures, and extending this bias into technological systems further exacerbates their marginalization. For example, in applications such as road or public safety, failure to detect or misclassification of these sections can have dire consequences, such as exclusion from safety measures or unequal access to services.

Our motivation comes from the ethical responsibility to address these systemic challenges, ensures representation of individuals in the development of AI, and foster a technological ecosystem that values diversity and inclusion.

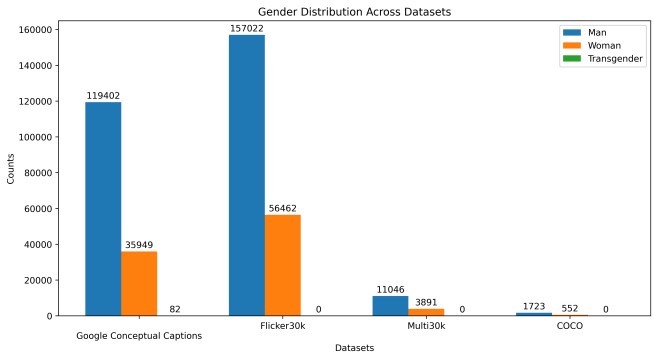

Fig. 1. Distribution across data sets. The chart highlights the counts of 'Man', 'Woman', and 'Transgender' mentions in four popular image captioning data sets.

A. Problem Statement and Scope of Work

In recent years, advances in AI have been driven by large-scale data sets that serve as the foundation for training and evaluating AI systems. However, a closer examination of these data sets reveals a significant disparity in representation, particularly for transgender individuals. This underrepresentation poses critical challenges in developing inclusive AI systems that serve all users equitably. The current scope of the work is limited to individuals who have declared their identity as transgender individuals. The focus is exclusively on visual datasets and does not address issues and biases in text-based or multimodal systems.

B. Data Set Analysis: Disparities in Representation

To illustrate this issue, we analyze four widely used data sets–Google's Conceptual Captions, Flickr30k, Multi30k, and COCO with respect to their representation of gender categories (see Fig. 1).

- **Conceptual Captions**: With a total of $2,007,090$ images, the data set includes annotations for $119,402$ images captioned as *man*, $35,949$ as *woman*, and only 82 as *transgender*.
- **Flickr30k**: Among $401,717$ images, $157,022$ are captioned as *man*, 56,462 as *woman*, and none as *transgender*.
- **Multi30k**: Among $29,000$ images, $11,046$ are captioned as *man*, $3,891$ as *woman*, and none as *transgender*.
- **COCO**: Despite its large size of $566,747$ images, only $1,723$ are captioned as *man*, 552 as *woman*, and none as *transgender*.

The above statistics highlight a troubling trend *viz.* transgender individuals are minimally represented or completely absent from these data sets. This skewed distribution fails to capture the diversity of human identity and experiences, leading to AI systems that inadvertently reinforce binary gender norms and marginalize these communities. In addition, the complex ethical challenges of collecting data, including privacy and potential misuse is rarely addressed.

C. Implications of the Data Gap

The lack of a spectrum of gender representation in image data sets has far-reaching implications.

1) **Bias in Model Predictions**: AI models trained on these data sets are likely to exhibit biased performance, misclassifying or entirely ignoring the marginalized individuals.
2) **Reinforcement of Binary Norms**: The exclusion of any of the marginalized labels perpetuates the binary understanding of gender, ignoring the spectrum of gender and fluidity of gender identities.
3) **Safety and Accessibility Concerns**: AI systems, such as facial recognition or action detection, may not serve everyone, leading to safety risks and reduced accessibility.

D. Related Work

Recent studies underscore the intersection of AI and transgender issues, emphasizing the importance of inclusion and ethical considerations. Bragazzi *et al.* [1] reviewed the impact of generative AI on the LGBTQ community, identifying both opportunities and challenges. Bias in AI technologies, particularly facial recognition, remains a critical concern. Perilo and Valença [2] conducted a systematic study on the limitations of facial recognition technologies for transgender individuals, revealing significant inaccuracies. Tripathi and Dubey [3] proposed safeguards to mitigate biases, ensuring better inclusion of transgender individuals in AI systems.

Sheridan [4] critically examined how AI narratives construct transgender identities, highlighting the social implications of biased datasets. Crabtree *et al.* [5] explored the potential for AI-driven persuasion to reduce prejudice and support transgender rights on a large scale.

The role of fairness in machine learning has also been studied extensively. Buolamwini and Gebru [6] highlighted accuracy disparities in commercial gender classification systems, while Keyes [7] discussed the challenges of automatic gender recognition for transgender and non-binary individuals. Broader ethical frameworks in AI were discussed by Binns [8], who examined lessons from political philosophy, and Feeney [9], who analyzed AI's societal impacts on the transgender community.

These studies collectively emphasize the necessity for diverse and inclusive datasets, fairness-aware algorithms, and ethical AI practices to address the biases and limitations affecting transgender representation in AI systems.

II. Proposed Approach

Addressing gender representation disparities in AI systems necessitates a comprehensive strategy:

- **Expanding Data Set Annotations**: Existing datasets should be revised to include diverse gender identities through self-declared samples rather than visual annotations, reducing the risk of bias and stereotyping.
- **Inclusive Data Collection**: Future data collection efforts must actively involve marginalized communities, ensuring their representation is accurate and consensual.
- **Establishing Ethical Frameworks**: Researchers should prioritize privacy, informed consent, and respectful representation by adhering to robust ethical guidelines throughout the data curation process.

By addressing these challenges, we aim to guide the development of AI systems that respect and reflect human diversity, fostering inclusion and equity in technological innovation.

A. Possible Applications

Incorporating the spectrum of gender representation in Generative AI enhances inclusion and addresses biases across critical domains-

- **Facial Recognition and Authentication:** Reducing biases in security systems for accurate identification.
- **Virtual Assistants:** Creating empathetic conversational agents with inclusive interactions.
- **Healthcare:** Supporting personalized care, including hormone therapy and mental health applications.
- **Media and Content Creation:** Generating authentic representation in media and entertainment.
- **Education and Awareness:** Promoting understanding through AI-generated learning materials.

B. Limitations in State-of-the-Art Approaches

Despite significant advances in AI, state-of-the-art approaches reinforces gender biases. Models such as Vision Transformers (ViT), ResNet, and YOLO, which are trained on biased data sets, often fail to classify, detect, and segment the marginalized individuals accurately. This is mainly due to the skewed or absent representation in existing data sets, leading to high misclassification rates and inconsistent performance.

Another critical limitation is the lack of ethical considerations and inclusive practices during data collection and

model training. This limitation perpetuates biases in downstream applications, further widening the gap in AI fairness. Although this study focuses on challenges within computer vision, it is essential to note that similar biases exist in text-based data sets, affecting natural language processing tasks such as sentiment analysis, language modeling, and chatbot interactions. However, addressing these biases in text data is beyond the scope of this article warrants further exploration in future research. These limitations underscore the pressing need for inclusive data sets and frameworks to ensure fairness and equity in AI systems.

C. Proposed Solutions to Address Limitations

To address the limitations and biases in AI systems concerning transgender representation, the following solutions are proposed:

1) **Creation of Inclusive Data Sets**: Developing ethically sourced, large-scale datasets with substantial representation of transgender individuals. Such datasets should span diverse applications, including pose estimation, facial recognition, and object detection, while ensuring informed consent and privacy protections.
2) **Self-Declaration-Based Data Collection**: Incorporating human-in-the-loop self-declaration mechanisms during data collection ensures that individuals can accurately represent their identities, reducing misrepresentation and annotation biases.
3) **Fairness-Aware Training Techniques**: Employing fairness-aware algorithms, such as adversarial debiasing and re-weighted loss functions, can help mitigate performance disparities across demographic groups and ensure equitable outcomes.
4) **Active and Continual Learning**: Leveraging active learning frameworks prioritizes the inclusion of underrepresented classes through iterative dataset refinement. Continual learning enables models to adapt to new data while preserving previously learned knowledge, promoting scalability and long-term inclusivity.
5) **Integration of Ethical AI Practices**: Embedding ethical guidelines across the AI development lifecycle, from dataset creation to model deployment, is essential. Collaborating with transgender and other marginalized communities during design, evaluation, and deployment phases provides valuable perspectives and ensures respectful representation.

By adopting these strategies, AI systems can become more inclusive, equitable, and robust, ultimately fostering fairness and enhancing their applicability in real-world scenarios.

D. Open-Source Data Set

To promote equitable AI, we are developing a large-scale dataset featuring images that include marginalized individuals. This dataset is designed to improve the performance of computer vision models in classification, detection, and segmentation tasks, thereby fostering greater inclusion and fairness in AI applications.

While the dataset is currently private to address privacy and ethical concerns, we plan to provide controlled access to researchers and practitioners who align with our commitment to ethical AI development. This approach ensures the protection of individual privacy while encouraging meaningful collaboration and the development of unbiased and robust models.

For further information and access requests, the dataset is hosted on Kaggle at: https://www.kaggle.com/datasets/pallabm22/genderspectrum-dataset.

III. Limitations

This work acknowledges several limitations:

1) **Focus on Visual Datasets:** Our study is limited to visual datasets, with multimodal datasets left as future work.
2) **Ethical Challenges:** Collecting data on underrepresented individuals involves complexities such as informed consent, privacy, and potential misuse. Any released datasets will be secured and accessible only through authentication mechanisms.
3) **Risk of Stereotypes:** Gender categorization in datasets risks reinforcing stereotypes, highlighting the need for inclusive and sensitive annotation practices.
4) **Beyond Datasets:** Addressing gender gaps requires attention not only to data but also to algorithmic design and societal factors, which fall outside the scope of this study.

These limitations underscore the complexity of the issue and the need for collaborative efforts to advance inclusive AI development.

IV. Future Work

In future work, we plan to address the limitations identified in this study. While this paper focuses on visual datasets, we aim to expand our scope to include multimodal datasets that integrate text, audio, and other modalities. This will enable a more comprehensive analysis of biases and fairness across diverse AI applications.

Additionally, we will fine-tune state-of-the-art models such as Vision Transformers (ViT), ResNet, YOLO, and U-Net using carefully curated datasets, emphasizing ethical considerations, informed consent, and privacy. Our objective is to address classification and segmentation gaps without reinforcing harmful stereotypes, ensuring better representation across gender identities in computer vision tasks.

Beyond datasets, we will explore algorithmic architectures and design strategies to mitigate biases inherent in AI systems. Furthermore, we aim to extend our efforts to evaluating fairness in text and multimodal data for large language models, addressing biases that transcend computer vision. By adopting a holistic approach, we seek to foster fairness, inclusion, and ethical AI development across multiple domains.

V. Conclusion

This article underscores the critical underrepresentation of marginalized individuals, particularly in terms of gender

diversity, in widely used computer vision datasets. Such gaps perpetuate biases in state-of-the-art AI systems, limiting their fairness and inclusivity. Our analysis reveals that models like Vision Transformers (ViT), ResNet, and YOLO struggle to reliably classify, detect, and segment individuals across the gender spectrum due to these dataset limitations.

To address this pressing issue, we propose the creation of a large-scale, inclusive dataset with a specific focus on transgender and gender-diverse representation. This initiative emphasizes ethical data collection practices, including self-declared identities through human-in-the-loop methodologies, and incorporates advanced learning techniques such as fairness-aware algorithms and active learning.

By adopting the proposed framework, we aim to bridge the data gap and enhance the fairness, accuracy, and inclusivity of AI systems. This work advocates for equitable representation in AI, particularly for historically marginalized communities, and lays out a roadmap for ethical and inclusive AI development. The open-sourcing of the proposed dataset ensures accessibility, fostering global collaboration and innovation in building equitable AI systems.

In conclusion, this study not only highlights the significance of addressing gender biases in AI but also provides actionable strategies to promote diversity, equity, and fairness, ultimately driving the development of AI systems that reflect and respect the diversity of human experiences.

REFERENCES

[1] N. L. Bragazzi, A. Crapanzano, M. Converti, R. Zerbetto, and R. Khamisy-Farah, "The impact of generative conversational artificial intelligence on the lesbian, gay, bisexual, transgender, and queer community: scoping review," *Journal of Medical Internet Research*, vol. 25, p. e52091, 2023.

[2] M. Perilo and G. Valença, "How facial recognition technologies affect the transgender community? a systematic mapping study," in *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, 2024, pp. 1153–1160.

[3] A. Tripathi and P. Dubey, "Breaking boundaries: Ai safeguards for women and transgender individuals," in *Developing AI, IoT and Cloud Computing-based Tools and Applications for Women's Safety*. Chapman and Hall/CRC, pp. 206–220.

[4] T. Sheridan, "Reading transness in ai narratives: How artificiality constructs transgender identity," Master's thesis, Florida Atlantic University, 2023.

[5] C. Crabtree, J. Holbein, M. Bosley, and S. Sevi, "Can ai reduce prejudice at scale? evaluating the effectiveness of ai-powered personalized persuasion on support for transgender rights," 2024.

[6] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[7] O. Keyes, "The misgendering machines: Trans/hci implications of automatic gender recognition," *Proceedings of the ACM on human-computer interaction*, vol. 2, no. CSCW, pp. 1–22, 2018.

[8] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 149–159.

[9] M. Feeney, "When ai meets the transgender community," 2022.