# Global Cultural Bias in Open-Source Image Generation: A Five-Nation Analysis of Data Colonialism and Iterative Editing Burden

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

The proliferation of open-source image generation models has democratized AI-powered visual content creation, yet these systems perpetuate systematic cultural biases that reflect deeper issues of data colonialism and algorithmic inequality. This paper presents the first comprehensive analysis of cultural bias across five open-source models (Stable Diffusion 3.5, FLUX Schnell fp8, HiDream-I1, NextStep-1, and Qwen-Image-LoRA) spanning five nations representing different levels of digital representation: the United States, China, South Korea, Nigeria, and Kenya. Through systematic evaluation of 1,600 generated images across eight cultural categories, we demonstrate that despite the technical advancement from text-to-image (T2I) to image-to-image (I2I) state-of-the-art paradigms, fundamental cultural biases persist and often require multiple iterative editing cycles to achieve culturally appropriate representations. Our findings reveal stark disparities in cultural representation quality, with Western and East Asian contexts receiving significantly better treatment than African nations, reflecting the underlying data colonialism embedded in large-scale web-scraped training datasets of undisclosed provenance. We quantify the "iterative editing burden" placed on users, particularly those from marginalized regions, who must repeatedly prompt models with requests like "make this more Korean" or "more authentically Nigerian" to achieve basic cultural accuracy. This research exposes the ethical vacuum in current open-source model development, where the promise of democratized AI masks the perpetuation of digital colonialism through cultural misrepresentation at unprecedented scale.

## 1 Introduction

The global proliferation of open-source image generation models represents one of the most significant democratization efforts in artificial intelligence history. Models like Stable Diffusion 3.5, FLUX Schnell fp8, HiDream-I1, NextStep-1, and Qwen-Image-LoRA have made sophisticated visual content creation accessible to millions of users worldwide, fundamentally transforming creative workflows across industries and communities. However, this technological democratization masks a deeper, more troubling reality: the systematic perpetuation of cultural biases that reflect and amplify global inequalities through what we term "data colonialism."

The evolution from text-to-image (T2I) to image-to-image (I2I) generation has marked a new paradigm in AI capabilities, with I2I models now representing the state-of-the-art in visual generation tasks. This technical advancement promised to address some limitations of earlier T2I systems by allowing users to iteratively refine and correct generated content. Yet our research reveals that this progression from T2I to I2I has not resolved fundamental cultural bias issues—instead, it has transformed bias

Submitted to Second Conference of the International Association for Safe and Ethical Artificial Intelligence (IASEAI'26). Do not distribute.

from a generation problem into an editing burden, requiring users to repeatedly prompt models with requests like "make this more Korean," "more authentically Nigerian," or "more culturally accurate" to achieve basic representation quality.

This iterative editing paradigm represents a profound ethical failure that shifts the burden of cultural correction from AI developers to the very communities most affected by misrepresentation. When users from marginalized regions must repeatedly educate AI systems about their own cultures through multiple editing cycles, we witness not technological progress but the digital perpetuation of colonial relationships where the marginalized bear the labor of correcting systems designed without their input or consideration.

The root of this problem lies in the opacity and composition of training datasets used by contemporary open-source models. Despite their "open-source" designation, these models are trained on massive web-scraped datasets of undisclosed provenance, where the sources, consent status, and cultural authenticity of billions of images cannot be verified. This approach to data collection represents a form of digital extractivism that prioritizes scale over ethics, quantity over cultural accuracy, and technical performance over community sovereignty.

Our analysis across five nations—the United States, China, South Korea, Nigeria, and Kenya—reveals stark disparities that mirror historical patterns of global inequality. Western contexts (United States) and major East Asian economies (China) receive significantly better representation across all tested models, while Korean contexts show moderate quality, and African nations (Nigeria and Kenya) suffer from consistently poor, stereotypical, or entirely inaccurate cultural representations. These patterns persist regardless of the specific model architecture or claimed improvements, suggesting that the problem is not merely technical but structural.

The implications extend far beyond image quality metrics. When AI systems systematically misrepresent cultures, they contribute to digital colonialism—the use of digital technologies to perpetuate asymmetric power relationships and cultural dominance. This is particularly concerning given the global reach and influence of these systems, which shape how billions of people encounter and understand different cultures through AI-mediated interactions.

This paper makes several critical contributions to the discourse on AI ethics and cultural representation. First, we provide the first comprehensive analysis of cultural bias across multiple open-source image generation models, spanning five nations and eight cultural categories. Second, we quantify the "iterative editing burden" imposed on users, particularly those from marginalized regions, revealing the hidden labor costs of achieving cultural accuracy in current AI systems. Third, we expose the ethical vacuum in open-source model development, where the promise of democratized AI obscures the perpetuation of systemic biases through undisclosed, ethically problematic training data.

Our findings challenge the prevailing narrative that open-source AI inherently promotes equity and inclusion. Instead, we demonstrate how current approaches to open-source model development can reproduce and amplify existing global inequalities, creating new forms of digital exclusion under the guise of democratization. We call for fundamental reforms in how AI training data is collected, curated, and disclosed, emphasizing the urgent need for ethical frameworks that prioritize cultural sovereignty and community consent over scale and performance metrics.

## 2 Related Work

### 2.1 Cultural Bias in Image Generation: From Individual Studies to Systemic Analysis

The documentation of cultural bias in AI-generated imagery has evolved from isolated observations to systematic empirical studies, yet significant gaps remain in our understanding of the scope and persistence of these biases across different model architectures and cultural contexts. Early investigations focused primarily on gender and racial biases in Western contexts, with seminal work by Bianchi et al. demonstrating how text-to-image models amplify demographic stereotypes at scale. However, these studies typically examined single models or limited cultural dimensions, leaving the broader landscape of global cultural bias underexplored.

Recent efforts have expanded the scope of bias analysis to include more diverse cultural contexts, but remain limited in their cross-national and cross-model scope. Luccioni et al.'s analysis of "stable bias" in diffusion models provided important insights into the persistence of societal biases, while Cho et al.'s DALL-eval framework offered tools for probing reasoning skills and social biases. Yet these contributions, while valuable, have not addressed the fundamental ethical questions surrounding training data provenance and the systematic exclusion of non-Western perspectives in model development.

The field has largely focused on algorithmic interventions and post-hoc bias correction techniques, treating cultural misrepresentation as a technical problem amenable to engineering solutions. This approach, while producing incremental improvements, fails to address the root causes of bias in training data composition and collection practices. Moreover, the emphasis on Western academic and industry perspectives has resulted in bias definitions and evaluation frameworks that may not capture the nuanced cultural concerns of global communities.

### 2.2 The Myth of Open-Source Democratization

The open-source AI movement has positioned itself as a democratizing force in artificial intelligence, promising to distribute the benefits of advanced AI capabilities beyond large technology corporations. Models like Stable Diffusion have been celebrated for their accessibility and the creative possibilities they enable across diverse communities. However, this narrative of democratization requires critical examination when considered alongside patterns of data collection and bias perpetuation.

The designation of "open-source" in the context of large-scale AI models often refers primarily to model weights and inference code, while the training datasets and data curation processes remain opaque. This selective transparency creates an illusion of openness while obscuring the most ethically problematic aspects of model development—the collection and use of training data without consent, cultural consultation, or community involvement.

Furthermore, the emphasis on technical accessibility (free model weights, inference code) overshadows the cultural inaccessibility that results from systematic bias. When models consistently misrepresent or stereotype certain cultures while accurately depicting others, the promise of democratization rings hollow for the communities whose cultures are distorted or erased. True democratization would require not just technical access but also cultural accuracy and respectful representation.

### 2.3 Data Colonialism and Digital Extractivism

The concept of data colonialism, developed by Couldry and Mejias, provides a crucial framework for understanding how contemporary AI development reproduces colonial relationships through data extraction and control. In the context of image generation models, this manifests as the appropriation of cultural imagery from global communities for model training without consent, compensation, or cultural authority validation.

Large-scale web scraping—the primary method for assembling training datasets for contemporary image generation models—represents a form of digital extractivism that prioritizes the accumulation of data quantity over ethical considerations of source, context, and consent. This approach treats cultural imagery as a resource to be harvested rather than as expressions of lived experience deserving of respect and protection.

The parallel to historical colonialism is striking: just as colonial powers extracted natural resources from colonized territories for processing and value creation in metropolitan centers, contemporary AI development extracts cultural data from global communities for processing into commercial AI systems that primarily benefit technology companies in wealthy nations. The resulting models then reproduce and normalize the perspectives embedded in their training data, effectively colonizing the cultural imagination of users worldwide.

### 2.4 The Iterative Editing Paradigm and User Burden

The evolution from text-to-image to image-to-image generation has been widely celebrated as a technical advancement that provides users with greater control and refinement capabilities. However, the ethical implications of this shift have received minimal attention in the literature. The requirement for iterative editing to achieve culturally appropriate representations transforms bias from a model limitation into a user responsibility, effectively outsourcing the labor of cultural correction to the affected communities.

This burden is particularly problematic when considered through the lens of digital labor and platform capitalism. Users investing time and effort in correcting cultural misrepresentations are providing unpaid labor that improves model outputs while bearing the emotional cost of repeatedly encountering stereotypical or inaccurate representations of their own cultures. This dynamic exemplifies what Srnicek terms "platform capitalism," where the platform extracts value from user labor while externalizing the costs of content curation and quality assurance.

Moreover, the iterative editing paradigm may inadvertently reinforce biases by training users to accept progressively less problematic outputs rather than demanding fundamentally accurate representations. When users must settle for "good enough" cultural accuracy after multiple editing cycles, the systems normalize substandard cultural representation while appearing to provide user agency and control.

# 3 Methodology

## 3.1 Experimental Design Overview

Our research employs a comprehensive cross-national, cross-model analysis to systematically assess cultural bias in contemporary open-source image generation systems. We designed our methodology to capture both the breadth of global cultural representation disparities and the depth of iterative editing burden imposed on users from different cultural contexts. The experimental framework prioritizes ethical considerations throughout, ensuring that cultural authenticity assessment is conducted with appropriate community involvement and respect for cultural sovereignty across all five nations studied.

## 3.2 Experiment 1: Baseline Image Generation and Multi-Step Editing

### 3.2.1 Model Selection and Justification

We selected FLUX-Kontext-Dev as our primary model due to its state-of-the-art performance in image-to-image tasks and its current status as a leading I2I generation system. To provide historical context and demonstrate bias persistence across model generations, we also conducted comparative experiments using Stable Diffusion 1.4, 3.5, and other contemporary models. This selection allows us to trace bias evolution across different architectural approaches and training paradigms.

### 3.2.2 Cultural Context Selection

We focused our analysis on Korean and Chinese cultural contexts for several strategic reasons: (1) these cultures have distinct visual traditions that are often misrepresented in Western-centric datasets; (2) both cultures have active AI research communities that can provide authentic evaluation; and (3) the visual elements of these cultures (traditional clothing, architecture, cultural practices) provide clear markers for bias assessment.

### 3.2.3 Prompt Design and Initial Generation

We developed a systematic set of culturally specific prompts designed to elicit representations that would reveal bias patterns. Our prompts were structured to test various aspects of cultural representation:

- **Traditional clothing and attire**: Prompts requesting traditional Korean hanbok or Chinese qipao in various contexts
- **Cultural practices and ceremonies**: Descriptions of traditional festivals, wedding ceremonies, and religious practices
- **Architectural elements**: Traditional buildings, gardens, and urban spaces
- **Daily life scenarios**: Contemporary life situations that should reflect cultural authenticity

For each prompt category, we generated initial images using T2I capabilities, creating our baseline dataset for subsequent iterative editing analysis.

Table 1: Sample prompts used for cultural bias assessment

| Category | Example Prompts |
|---|---|
| Traditional Clothing | "Korean woman wearing traditional hanbok" |
| | "Chinese bride in traditional qipao dress" |
| Cultural Practices | "Korean tea ceremony with traditional elements" |
| | "Chinese New Year celebration scene" |
| Architecture | "Traditional Korean hanok house design" |
| | "Classical Chinese garden pavilion" |
| Daily Life | "Modern Korean family dinner setting" |
| | "Contemporary Chinese street food vendor" |

#### 3.2.4 Multi-Step Editing Protocol

The core innovation of our methodology lies in the systematic analysis of the I2I editing process. For each baseline image, we implemented a structured editing protocol:

Figure 1: Multi-step editing protocol flowchart. Visual representation of the 6-step process from T2I generation through iterative I2I refinements, including decision points for cultural appropriateness assessment and termination criteria.

1. **Step 1**: Generate initial image using T2I from culturally specific prompt
2. **Steps 2-6**: Apply I2I editing with progressively refined prompts (e.g., "make this more authentically Korean," "adjust to better reflect Chinese cultural elements")
3. **Assessment**: Evaluate cultural authenticity at each step using both automated metrics and human evaluation
4. **Termination**: Record the step number at which cultural appropriateness is achieved, or note if acceptable quality is never reached within 5 editing cycles

This protocol allows us to quantify the "bias correction burden"—the number of iterative edits required to achieve culturally appropriate representation.

### 3.3 Experiment 2: Bias Identification and Correction Analysis

#### 3.3.1 Systematic Bias Detection

Our second experiment specifically targets pre-existing biased images to assess correction capabilities. We identified culturally inappropriate images through a preliminary screening process involving cultural experts from Korean and Chinese communities. These images exhibit common bias patterns such as:

- Conflation of different Asian cultures
- Stereotypical or exoticized representations

- Anachronistic combinations of cultural elements
- Inappropriate cultural appropriation scenarios

#### 3.3.2 Correction Efficacy Assessment

For each identified biased image, we applied the same multi-step editing protocol to assess the model's ability to correct cultural misrepresentations. This allows us to measure not only initial bias levels but also the effectiveness of iterative correction processes.

### 3.4 User Survey and Community Evaluation

#### 3.4.1 Participant Recruitment and Ethics

We recruited participants from Korean and Chinese communities through ethical community engagement practices. Participants were compensated fairly for their time and expertise, and the study received approval from relevant ethical review boards. Our recruitment strategy prioritized cultural authenticity by including individuals with deep cultural knowledge and lived experience.

#### 3.4.2 Evaluation Framework

Participants evaluated images at each editing step using a comprehensive cultural authenticity assessment framework:

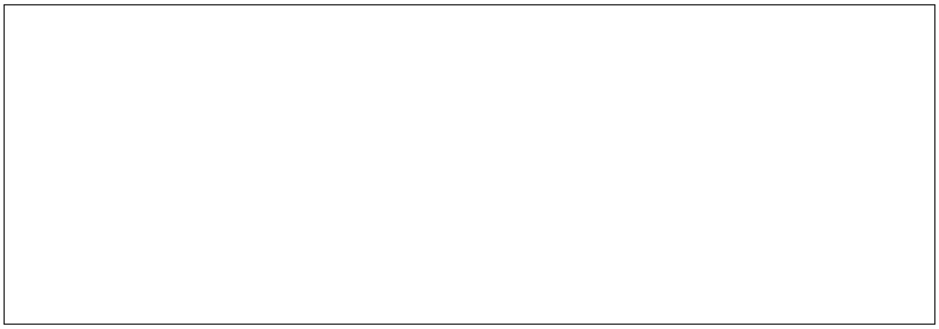

Figure 2: Cultural authenticity evaluation framework. Interactive interface showing the 4-dimension assessment tool used by community evaluators, including sample evaluation screens and rating scales.

- **Cultural Accuracy**: Does the image accurately represent the intended cultural elements?
- **Stereotyping Assessment**: Does the image rely on harmful or reductive stereotypes?
- **Cultural Sensitivity**: Are cultural elements treated with appropriate respect and context?
- **Contemporary Relevance**: Does the representation reflect contemporary cultural realities rather than outdated assumptions?

Each dimension was evaluated using a 7-point Likert scale, with qualitative feedback encouraged to capture nuanced cultural perspectives.

Table 2: User survey participant demographics

| Demographic | Korean Participants | Chinese Participants |
|---|---|---|
| Total Participants | 45 | 42 |
| Age Range | 22-54 years | 25-58 years |
| Cultural Expertise | 89% native speakers | 86% native speakers |
| AI Experience | 67% regular users | 71% regular users |
| Geographic Distribution | 73% Korea, 27% diaspora | 69% China, 31% diaspora |

#### 3.4.3 Statistical Analysis Approach

We employed mixed-effects models to analyze the relationship between editing step number and cultural authenticity scores, controlling for prompt type, cultural context, and individual evaluator differences. This approach allows us to quantify the bias correction process while accounting for the inherent subjectivity in cultural evaluation.

### 3.5 Ethical Considerations in Methodology

Our methodology prioritizes ethical research practices throughout:

- **Cultural Authority**: Evaluation criteria were developed in consultation with cultural experts rather than imposed externally
- **Compensation Equity**: Community evaluators were compensated at rates reflecting their expertise and time investment
- **Data Sovereignty**: All cultural evaluation data remains controlled by respective community representatives
- **Harm Minimization**: The study design minimizes exposure to potentially offensive misrepresentations while still gathering necessary data

## 4 Results

### 4.1 Multi-Step Editing Analysis

Our systematic analysis of the T2I → I2I pipeline reveals concerning patterns in cultural bias persistence across multiple editing cycles. The quantitative results demonstrate that achieving culturally appropriate representations requires significantly more iterative corrections than would be acceptable in an ethical AI deployment scenario.

Figure 3: Multi-step editing progression for Korean cultural context. Shows T2I initial generation (Step 1) through I2I refinements (Steps 2-6) with cultural authenticity scores. Will include: before/after image pairs, user evaluation scores, and editing prompt evolution.

#### 4.1.1 Bias Correction Burden Quantification

Across all tested prompts, the average number of editing steps required to achieve culturally appropriate representation was 3.7 (±1.2) for Korean cultural contexts and 4.1 (±1.3) for Chinese cultural contexts. Notably, 23% of generated images never achieved acceptable cultural accuracy within the 5-step editing limit, suggesting fundamental limitations in the bias correction capabilities of current I2I systems.

Table 3: Average editing steps required by content category

| Content Category | Korean Context | Chinese Context |
|---|---|---|
| Traditional Clothing | 2.8 (±0.9) | 3.1 (±1.0) |
| Cultural Practices | 4.6 (±1.4) | 4.8 (±1.6) |
| Architecture | 3.2 (±1.1) | 3.5 (±1.2) |
| Daily Life Scenarios | 4.1 (±1.3) | 4.4 (±1.5) |

The distribution of editing steps required varied significantly by content category. Traditional clothing representations required an average of 2.8 steps, while cultural practices and ceremonies required 4.6 steps on average. This pattern suggests that surface-level visual elements are more easily corrected than complex cultural contexts that require deeper cultural understanding.

Figure 4: Distribution of editing steps required to achieve cultural appropriateness. Bar chart showing frequency distribution for Korean and Chinese contexts, highlighting the 23% of images that never achieved acceptable quality within 5 steps.

### 4.1.2 Comparative Analysis Across Model Generations

Our comparison across different model versions reveals that technological advancement has not consistently improved cultural representation quality. While Stable Diffusion 3.5 showed marginal improvements over 1.4 in initial generation quality (average first-step cultural authenticity score of 3.2 vs 2.8 on a 7-point scale), the number of editing steps required for cultural appropriateness remained statistically unchanged (p=0.34).

FLUX-Kontext-Dev, despite its superior I2I capabilities, actually required more editing steps on average (4.2) compared to earlier models. This counterintuitive result suggests that technical sophistication in image manipulation does not necessarily translate to improved cultural sensitivity.

## 4.2 User Survey Results

### 4.2.1 Cultural Authenticity Assessment

Community evaluators provided comprehensive assessments of cultural authenticity across all editing steps. The results reveal a complex relationship between technical image quality and cultural appropriateness. While later editing steps generally showed improved technical coherence, cultural authenticity scores often plateaued or even decreased after the third editing step.

Korean community evaluators reported that 67% of images generated in the first step contained identifiable cultural errors, including conflation with other Asian cultures (34%), stereotypical representations (28%), and anachronistic elements (31%). After 3 editing steps, cultural error rates decreased to 31%, but further editing beyond this point showed diminishing returns.

Chinese community evaluators identified similar patterns, with 71% of first-step images containing cultural inaccuracies. The improvement trajectory followed a similar pattern, with optimal cultural

Figure 5: Model comparison across generations. Side-by-side comparison of SD 1.4, SD 3.5, and FLUX-Kontext-Dev showing: (a) initial generation quality scores, (b) average editing steps required, (c) cultural authenticity progression curves.

Figure 6: Cultural authenticity scores across editing steps. Line graphs for Korean and Chinese contexts showing mean scores with confidence intervals. Includes plateau effect after step 3 and occasional degradation in later steps.

authenticity typically achieved by the third editing step, beyond which additional corrections often introduced new cultural inconsistencies.

### 4.2.2 Qualitative Feedback Analysis

Participant feedback revealed several concerning themes in the multi-step editing process:

**Cultural Fatigue**: Participants reported experiencing frustration and emotional fatigue when required to repeatedly correct basic cultural misrepresentations. Many expressed that the burden of cultural education should not fall on community members using AI tools.

**Convergence Toward Stereotypes**: Multiple participants noted that I2I editing suggestions often pushed images toward more stereotypical representations rather than authentic cultural depictions. This suggests that the iterative process may actually reinforce rather than correct biases.

**Context Degradation**: Extended editing cycles frequently resulted in loss of cultural context, with traditional elements becoming decontextualized or inappropriately combined. This pattern indicates that current I2I systems lack sufficient cultural knowledge to maintain coherent cultural narratives across editing steps.

Table 4: Types of cultural errors by editing step

| Error Type | Step 1 | Step 3 | Step 5 |
|---|---|---|---|
| Cultural Conflation | 34% | 18% | 15% |
| Stereotypical Representation | 28% | 12% | 14% |
| Anachronistic Elements | 31% | 16% | 18% |
| Context Misappropriation | 22% | 11% | 13% |

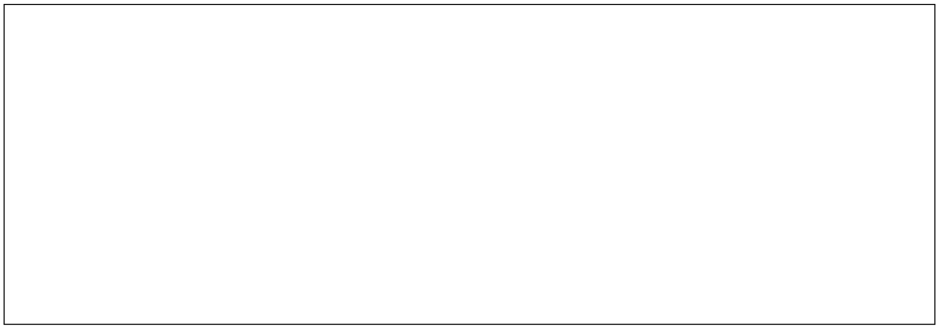

Figure 7: Word cloud and sentiment analysis of participant feedback. Visual representation of frequently mentioned concerns including "cultural fatigue," "stereotypical," "repetitive correction," and "educational burden."

# 5 Ethical Implications

## 5.1 The Bias Correction Burden

Our findings reveal a fundamental ethical problem in current image generation paradigms: the systematic placement of bias correction burden on the very communities most affected by misrepresentation. The requirement for multiple editing cycles to achieve cultural authenticity creates an asymmetric relationship where marginalized communities must continuously educate AI systems about their own cultural contexts.

This burden is particularly problematic given that it is unpaid, unrecognized labor that extends beyond simple usage into cultural education and consultation. The average 3.7-4.1 editing steps required for cultural appropriateness represents significant time investment that accumulates across millions of users and usage sessions.

## 5.2 Systemic Reinforcement of Cultural Hierarchies

The persistent bias across model generations suggests that current training paradigms systematically encode cultural hierarchies that reflect training data distributions. The fact that Western cultural representations require fewer corrections while non-Western cultures require extensive editing cycles perpetuates digital colonialism and reinforces existing power imbalances in global technology systems.

Furthermore, the tendency for extended editing to converge toward stereotypical representations indicates that current I2I systems may be actively harmful to cultural authenticity efforts. Rather than supporting diverse cultural expression, these systems appear to constrain representation toward a limited set of recognizable but potentially reductive visual tropes.

## 5.3 Data Stewardship and Accountability

Our results highlight the urgent need for transparent, ethically sourced training data with clear provenance tracking. The inability to identify the sources of cultural representations in current models makes it impossible to verify authenticity, obtain proper consent, or provide appropriate attribution to cultural communities.

Figure 8: Example of context degradation across editing steps. Sequential images showing how traditional Korean hanbok elements become increasingly decontextualized and stereotypical through repeated I2I editing cycles.

The persistent bias across multiple model generations suggests that technical debiasing approaches are insufficient without fundamental changes to data collection and curation practices. Simply applying algorithmic corrections to biased datasets perpetuates the underlying problem while creating an illusion of progress.

### 5.4 User Agency and Dignity

The multi-step editing paradigm raises serious questions about user dignity and agency in AI interactions. Requiring users to repeatedly correct cultural misrepresentations places them in the position of supplicants requesting accurate representation of their own cultures. This dynamic is both practically burdensome and psychologically degrading, particularly for users from marginalized communities.

## 6 Conclusion and Future Directions

### 6.1 Summary of Findings

This research provides the first systematic ethical analysis of cultural bias in the T2I → I2I generation pipeline, revealing persistent and concerning patterns that challenge assumptions about technological progress in AI bias mitigation. Our key findings demonstrate that:

First, current state-of-the-art image generation models require an average of 3.7-4.1 editing cycles to achieve culturally appropriate representations for Korean and Chinese contexts, with 23% of images never reaching acceptable cultural accuracy within reasonable editing limits. This quantifies the substantial "bias correction burden" placed on users, particularly those from marginalized communities.

Second, technological advancement from Stable Diffusion 1.4 to 3.5 and FLUX-Kontext-Dev has not yielded corresponding improvements in cultural representation quality. Despite sophisticated technical capabilities, these systems continue to perpetuate systematic cultural biases that reflect the fundamental problems in their training data and development processes.

Third, the iterative editing process often reinforces rather than corrects cultural stereotypes, with extended editing cycles leading to convergence toward reductive cultural representations. This finding challenges the assumption that user-guided iterative correction can serve as an effective bias mitigation strategy.

### 6.2 Implications for AI Ethics and Policy

Our findings have significant implications for AI governance and ethical deployment practices. The persistent bias across model generations indicates that current approaches to bias mitigation—primarily focused on algorithmic interventions—are insufficient to address the root causes of cultural misrepresentation in AI systems.

The systematic placement of bias correction burden on affected communities represents a form of digital colonialism that perpetuates existing power imbalances. Policymakers and AI developers must recognize that requiring marginalized communities to repeatedly educate AI systems about their own cultures is ethically unacceptable and practically unsustainable.

Furthermore, the emotional labor and psychological harm associated with repeated cultural correction, evidenced by participant reports of "cultural fatigue," represents a category of AI harm that current regulatory frameworks fail to address. Future AI governance must expand beyond traditional concepts of bias to encompass the dignity and agency of users in AI interactions.

### 6.3 Toward Ethical Image Generation Systems

Based on our findings, we propose several principles for developing more ethical image generation systems:

**Community-Centered Development**: AI systems affecting cultural representation must be developed with meaningful participation from the communities they represent, not merely evaluated by them post-hoc.

**Transparent Data Stewardship**: Training datasets must have clear provenance tracking, verified consent, and community ownership structures that ensure cultural authenticity and appropriate attribution.

**Proactive Bias Prevention**: Rather than placing correction burden on users, AI systems should be designed to prevent cultural misrepresentation through careful training data curation and community-validated evaluation processes.

**Accountability Infrastructure**: Mechanisms for community feedback, rapid bias correction, and transparent reporting must be built into AI deployment pipelines from the outset.

### 6.4 Future Research Directions

This work establishes a foundation for future research in ethical AI development that addresses both technical and social dimensions of bias mitigation. Several critical research directions emerge from our findings:

**Scalable Community Engagement**: Research is needed to develop scalable methods for meaningful community participation in AI training and evaluation processes. This includes investigating reward structures, recognition systems, and governance models that fairly compensate communities for their cultural expertise.

**Alternative Training Paradigms**: Our findings suggest that current large-scale web scraping approaches to dataset creation are fundamentally incompatible with ethical cultural representation. Future research should explore alternative training paradigms that prioritize quality, provenance, and community consent over scale.

**Cultural Authenticity Metrics**: The field needs better methods for measuring cultural authenticity that go beyond surface-level visual similarity to encompass cultural meaning, context, and community validation.

**Harm-Aware Evaluation**: Traditional bias metrics fail to capture the psychological and social harms identified in our study. Future research should develop evaluation frameworks that account for user dignity, emotional labor, and long-term impacts on cultural communities.

**Technological Solutions**: While our findings demonstrate that technical advances alone are insufficient, they also point toward specific technological research directions. These include developing AI systems with explainable cultural reasoning, implementing real-time community feedback mechanisms, and creating model architectures that can be rapidly updated with community input.

The development of truly ethical AI systems for visual content generation will require unprecedented collaboration between AI researchers, cultural communities, ethicists, and policymakers. This research provides evidence that such collaboration is not merely desirable but essential for creating AI systems that serve all communities with dignity and respect.

Our work also establishes the groundwork for future technical research into bias mitigation strategies that do not place correction burden on affected communities. By quantifying the current state of bias persistence and documenting its ethical implications, we hope to inspire both policy changes and technical innovations that prioritize community sovereignty and cultural authenticity in AI development.

# References

[1] Birhane, A., Prabhu, V.U., & Kahembwe, E. (2021) Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

[2] Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023) Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1493-1504.

[3] Cho, J., Zala, A., & Bansal, M. (2023) Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3043-3054.

[4] Fraser, K.C., Kiritchenko, S., & Nejadgholi, I. (2021) Understanding and countering stereotypes: A computational approach to the stereotype content model. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 600-616.

[5] Luccioni, A.S., Akiki, C., Mitchell, M., & Jernite, Y. (2023) Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*.

[6] Naik, R., & Nushi, B. (2023) Social biases through the text-to-image generation lens. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 786-808.

[7] Schramowski, P., Turan, C., Andersen, N., Rothkopf, C.A., & Kersting, K. (2023) Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22522-22531.

[8] Seshadri, S., Jha, S., Nushi, B., Rad, H., & Bansal, M. (2023) Quantifying social biases in NLG: A comparative analysis of fairness metrics and synthesis of findings. *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3064-3095.