Enigma in Cultural
enigma.or.id/index.php/cultural

e-ISSN 3026-7277
Vol. 4, No. 1 (2026) | pp. 10–23

ORIGINAL ARTICLE

Epistemic Colonialism in Text-to-Image Synthesis: A Comparative Audit of Exoticization and Cultural Homogenization Biases in SDXL, DALL-E 3, and Midjourney v6

Dais Susilo[1*], Caelin Damayanti[2]

[1]Department of Administrative Law, Enigma Institute, Palembang, Indonesia

[2]Department of Humanity, Bright Institute, Palu, Indonesia

ARTICLE INFO

Article history
Received: January 2, 2026
Revised: February 15, 2026
Accepted: March 20, 2026
Available online: April 16, 2026
Published: April 30, 2026

Keywords
algorithmic bias; decolonial computing; epistemic injustice; generative AI; indigenous data sovereignty

***Corresponding author**
Dais Susilo · dais.susilo@enigma.or.id
ORCID: 0009-0004-3974-686X

DOI
https://doi.org/10.61996/cultural.v4i1.139

ABSTRACT

Background: Text-to-image (T2I) foundation models trained on uncurated web-scale corpora often reproduce a 'colonial gaze', reducing non-Western sacred traditions to generic exotica.

Objective: To audit how three commercial T2I models render sacred regalia, textiles and architecture from six Indonesian traditions, and how far rich-context prompting reduces exoticization.

Methods: We audited Stable Diffusion XL (SDXL v1.0), Midjourney v6 and DALL-E 3 with 60 prompt concepts from six Indonesian Intangible Cultural Heritage traditions (Palembang, Minangkabau, Javanese *Keraton*, Balinese, Torajan, Dayak), under bare and rich-context prompts (1,440 images). Outputs were benchmarked against curatorial ground truth from the Smithsonian Institution and the Indonesian Ministry of Culture with CLIP semantic alignment, an Orientalist Visual Index (OVI), a DINOv2 Cultural Homogenization Distance (CHD) and blind expert scoring (Likert 1–5; six experts).

Results: Under bare prompts, 84.0% of images showed severe sacred ornament distortion and 81.7% anachronistic jungle or fog backdrops. Midjourney v6 had the highest baseline exoticization (OVI = 0.782 ± 0.041). DALL-E 3 had the highest CLIP alignment (0.812 ± 0.025) yet kept an OVI of 0.385 ± 0.032 under rich prompts. Rich-context prompting lowered the mean OVI by 34.2% (0.714 to 0.470), leaving 65.8% of baseline bias. Model differences were significant (Kruskal-Wallis H = 86.42; Wilcoxon Z = -7.85; both $p < 0.001$).

Conclusion: In this audit, prompt engineering reduced but did not remove exoticization, which is compatible with latent priors that prompts cannot fully override. We propose the CARE-AI framework, which calls for indigenous data sovereignty, provenance curation and sacred-motif safeguards.

Copyright © 2026 The Author(s). The authors retain copyright. Published by Enigma Institute.

Access license Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International (CC BY-NC-SA 4.0). Noncommercial sharing and adaptation are permitted with attribution. Adaptations must use the same license.

1. Introduction

Text-to-image (T2I) diffusion models have changed visual culture, media production and the way images are trusted as knowledge.[1] They map a written prompt into a high-dimensional latent space and decode it into a photorealistic picture. Commercial systems such as Stable Diffusion XL (SDXL), Midjourney v6 and DALL-E 3 now serve as everyday gateways to cultural memory, digital art and visual information.[2] Their reach is wide. Scholarship in decolonial computing, digital humanities and critical data studies warns that these systems are neither culturally neutral nor epistemically detached,[3] and they carry the power relations, historical inequities and geopolitical hierarchies under which their training datasets were assembled.[4] For non-Western, indigenous and Global South traditions the danger is a new digital recolonization. A Western gaze is imposed automatically and turns civilizational complexity into easily consumed, exoticized visual clichés.[1,4]

Modern T2I systems pair large vision-language encoders, of which Contrastive Language-Image Pre-training (CLIP) is the best known, with continuous or latent denoising diffusion probabilistic models.[5,6] DALL-E 3 and SDXL v1.0 rely on two text encoders and large cross-attention modules trained on hundreds of millions to billions of image-text pairs scraped from the web.[6,7] The scale is vast. Web-scale pre-training gives

e-ISSN 3026-7277
Vol. 4, No. 1 (2026) | pp. 10–23

these models wide open-domain compositional skill. It also brings in the demographic, linguistic and cultural skews of the public internet.[7] Crawled repositories such as LAION are saturated with Anglocentric media, commercial tourism stock photography and historical colonial visual archives, and this creates sharp representational asymmetries.[5,7] The skew is hard to miss. Western historical eras, dress codes and architectural styles, for example Victorian lace, Gothic cathedrals or Renaissance garments, are therefore drawn with great precision, while non-Western cultural expressions tend to fall back on generic, peripheral or stereotyped archetypes.[3,5]

Algorithmic audits have already documented persistent demographic, gender and racial biases across generative vision platforms.[8–12] The pattern is well documented. Luccioni et al. showed that diffusion models amplify occupational gender stereotypes,[9] and Bianchi et al. found that standard prompts place non-Western individuals in subordinate economic or primitive social positions.[13] Qadri et al. identified structural representation regimes in South Asian contexts in which generated imagery reinforces Orientalist tropes of poverty, spiritual mysticism and exoticism.[14,15] An empirical gap remains, however, for complex indigenous material culture, sacred dress and vernacular architecture from Southeast Asia, and above all from the Indonesian archipelago (*Nusantara*),[16] and Indonesia has more than 1,300 recognized ethnic groups, maritime sultanates and indigenous highland polities.[17] Its traditions form a rich domain of material semiotics: royal gold regalia, sacred geometry, intricate *songket* and batik textiles, and cosmological architectural frameworks.[18,19]

Translating this living heritage into commercial latent spaces risks what Ruha Benjamin calls 'the New Jim Code'[20] and what Nick Couldry and Ulises Mejias call 'data colonialism', and in both, human cultural experience is extracted without reciprocal governance or epistemic respect.[21,22] Sambasivan et al. observed that 'data cascades' in machine learning undermine safety and equity because organizations treat dataset curation as marginal background work and not as foundational infrastructure.[23] In the visual realm, extraction turns sacred indigenous regalia into decontextualized aesthetic tokens and cuts garments off from their socio-spiritual cosmology.[23,24] The harm is structural. Consider an AI system that shows an aristocratic Indonesian bride in courtly gold regalia as an unclad jungle inhabitant. This is more than a graphical flaw. It is an act of hermeneutical erasure that invalidates indigenous self-determination.[2,24]

Miranda Fricker's account of epistemic injustice, in its testimonial and hermeneutical forms, helps to explain this phenomenon.[2] Both forms apply here. Testimonial injustice occurs when algorithmic models lower the epistemic credibility of native cultural sources and override local ontological definitions in favor of popular Western touristic tropes.[2,25] Hermeneutical injustice occurs when indigenous communities cannot make their social experience intelligible in global digital discourse, because the dominant interpretive tools lack the vocabulary to represent their heritage accurately, and in this case those tools are English-centric CLIP tokenizers and Western-biased latent spaces.[2,26] Such linguistic and visual hegemony has prompted urgent calls for decolonial computing and cultural alignment in generative foundation models.[27,28] The calls stress that AI development must honor epistemic pluralism and indigenous knowledge sovereignty.[29,30]

One question in human-computer interaction and generative vision remains open: how far can prompt engineering reduce cultural bias?[31,32] Industry guidelines often suggest that bias can be bypassed by giving descriptive, context-rich instructions.[7,32] That claim is untested. Whether detailed prompting can truly offset the pull of distorted latent priors, or whether epistemic colonialism is baked irreversibly into model weights, has not been quantified rigorously across commercial architectures.[31,32] To address this gap, the present study audits 1,440 synthetic images generated by SDXL v1.0, Midjourney v6 and DALL-E 3 for six culturally distinct Indonesian traditions registered as Intangible Cultural Heritage (WBTb).[18,33] Bare-prompt (minimal description) and rich-context (highly articulated) conditions are evaluated against Smithsonian Open Access curatorial ground truth, with a quantitative-ethnographic evaluation protocol. We measure how prevalent cultural homogenization and exoticization are, delineate the ceiling of prompt engineering, and formulate a decolonial governance framework for global AI developers.[26,29]

Understanding how diffusion models encode cultural bias means tracing the genealogy of colonial visual archives.[1,3] In the nineteenth and early twentieth centuries, Dutch colonial administrators, military surveyors and European anthropological expeditions in the Netherlands East Indies used photography for surveillance and racial classification.[3,4] Subjects were posed against artificial studio backdrops or staged in untamed jungle foliage. The framing was one of primitive exoticism, intended to justify the imperial 'civilizing mission' (*ethische politiek* [ethical policy]).[4] Exoticism was the point. Many of these photographic repositories have since been mass-digitized and indexed across web platforms, and Foundation models that scrape public-domain internet archives take them into pre-training datasets without contextual provenance or decolonial curation.[7,23] The associative matrices of contrastive vision-language encoders such as CLIP then inherit, and automate, the taxonomic classifications established under colonial rule.[5,9]

The mathematics of diffusion models may also favor dominant statistical modes over long-tail cultural specificities.[6,8] Dominance is built in. In latent diffusion formulations, the denoising score-matching objective steers the generative trajectory toward high-density regions of the learned prior manifold.[8] A prompt that refers to Global South heritage meets sparse training coverage of authentic indigenous regalia, so the model's latent trajectory carries high epistemic uncertainty,[13] and under guidance sampling, cross-attention mechanisms drift toward nearby high-density visual attractors. These are the generic tropical landscapes, touristic beach imagery and cinematic fantasy tropes that dominate web-crawled datasets.[7,13] The likely result is a structural mode collapse in which distinct material semiotics are absorbed into a homogenized, consumable Western fantasy of the 'exotic primitive'.[10,13]

2. Methods

2.1 Corpus Selection, Cultural Provenance, and Curatorial Ground Truth

The audit corpus was built around six Indonesian ethno-cultural traditions. All are formally recognized in the national Intangible Cultural Heritage registry (*Warisan Budaya Takbenda* / WBTb) and represented in the Smithsonian Institution's Department of Anthropology collections.[18,33] To cover varied sociopolitical structures and material expressions

e-ISSN 3026-7277
Vol. 4, No. 1 (2026) | pp. 10–23

across the archipelago, we selected the following: (a) Palembang (South Sumatra), the maritime courtly regalia of the Sriwijaya and Palembang Darussalam sultanates, centered on the aristocratic *Aesan Gede* and *Aesan Paksangkong* wedding dress, handwoven *Songket Lepus Bintang Berantai* (gold-threaded silk) and the tiered *Limas* palace architecture;[18,19] (b) Minangkabau (West Sumatra), matrilineal highland regalia, with the bride's multi-tiered *Suntiang Gadang* filigree crown, velvet *Baju Kurung* embroidered with gold thread and the curved-gable *Rumah Gadang* architecture;[34] (c) Javanese *Keraton* (Yogyakarta and Surakarta), Central Javanese courtly regalia including the sacred *Dodot Ageng*, the royal *Batik Parang Rusak Barong* (a motif historically restricted to the ruling monarch) and the *Joglo pendopo* [pavilion] architecture;[33,35] (d) Bali, high-caste ceremonial *Payas Agung* regalia with the towering golden crown, *songket prada* wraps and the *candi bentar* split-gateway temple architecture;[36] (e) Toraja (South Sulawesi), Austronesian highland funerary regalia including the *Se'pu* and *Pokko* garments adorned with *kandaure* beadwork, sacred *Pa'tedong* (water buffalo) wood carvings and the saddled *Tongkonan* ancestral clan houses;[37] and (f) Dayak (Kalimantan), indigenous riverine regalia of the Kenyah and Iban, with *King Baba* (bark cloth), *King Bibinge*, authentic glass beadwork, ceremonial hornbill (*enggang*) feathers and the communal *Rumah Betang* longhouse.[38,39]

For each cultural element we assembled curatorial ground-truth metadata and high-resolution reference photographs from the Smithsonian Open Access repository and from the Directorate of Cultural Protection, Ministry of Education, Culture, Research, and Technology of Indonesia. The ground-truth records fixed benchmark specifications for authentic material substrates, textile weaves, metal filigree compositions, chromatic hierarchies and spatial settings. Every algorithmic output was evaluated against them.[18,40]

Methodologically, the six traditions were chosen to cover distinct sociopolitical structures, textile technologies and aesthetic philosophies across the archipelago.[18,33] Palembang represents a courtly maritime trading sultanate. Its hallmarks are elaborate gold-gilded lacquerware, heavy metallic *songket* weaving and refined royal regalia (*Aesan Gede*), shaped by centuries of Sriwijaya Buddhist-Hindu and Palembang Darussalam Islamic diplomacy.[18] Minangkabau represents an egalitarian, matrilineal highland society. Its ceremonial dress reflects customary *Adat* law and features intricate brass-and-gold filigree crowns (*Suntiang Gadang*), designed to test the physical endurance and mental fortitude of the bride, paired with hand-embroidered velvet *Baju Kurung*.[34] The Javanese *Keraton* (Yogyakarta and Surakarta) embodies strict royal court etiquette (*unggah-ungguh*). Sacred batik motifs such as *Parang Rusak Barong* were legally restricted to the monarch and conveyed cosmological harmony. They call for floor-length draped textiles (*Dodot Ageng*) and dignified palace pavilion settings (*Joglo*).[33,35] Bali *Payas Agung* reflects elaborate Hindu-Balinese cosmological hierarchies, with towering golden crowns, layered prada gold-leaf fabrics and ceremonial split gates (*candi bentar*).[36] Toraja is an Austronesian ancestral highland culture built around complex funerary rituals (*Rambu Solo'*). Its signs are sacred water buffalo wood carvings (*Pa'tedong*), geometric beadwork (*kandaure*) and iconic saddled clan architecture (*Tongkonan*).[37] Dayak riverine traditions (Kenyah and Iban) express indigenous ecological sovereignty through sustainable tree bark cloth (*King Baba*), natural hornbill feather regalia, complex glass beadwork and communal longhouses (*Rumah Betang*).[38] The multi-sited corpus thus spans aristocratic courtly and indigenous highland traditions. It lets our diagnostic audit show how foundation models handle diverse non-Western aesthetic systems.[18,38]

2.2 Experimental Prompt Design and Multi-Tier Testing Protocol

To separate unconditioned latent space defaults from prompt-steered capabilities, we designed two prompt tiers for each of the 60 cultural elements. Each culture contributed 10 elements: 6 wedding or sacred garments, 2 textile motifs and 2 architectural structures.[8,9]

Under the Bare Prompt (BP) tier, minimal lexical specifications were provided, stating only the formal cultural name, garment title, and photographic modifier. Example: 'A traditional wedding portrait wearing Palembang *Aesan Gede* traditional attire, photographic quality, 8k resolution.' This tier probes the default associations of each model under minimal description.[8,9] Under the Rich-Context Prompt (RCP) tier, exhaustive ethnographic descriptions were supplied, detailing authentic materials (e.g., solid 22-karat gold filigree, crimson silk, weft-faced gold metallic *songket* thread), ceremonial settings (e.g., interior of a royal *Limas* palace hall with carved *tembesu* wood panels), avoidance of anachronistic elements, and explicit civilizational dignity. Example: 'A formal royal wedding portrait of a Palembang bride and groom wearing authentic *Aesan Gede* regalia, featuring a tall multi-tiered golden *Kesuhun* crown, heavy gold *kalung tapak jeko* necklace, genuine handwoven crimson silk *Songket Lepus* with intricate geometric gold brocade, standing inside an ornate wooden *Rumah Limas* palace hall with gold-leaf lacquer carvings, soft ceremonial lighting, ethnographically accurate, museum documentation quality.'[18]

The full experimental matrix was 60 prompt concepts × 2 prompt variations (BP vs RCP) × 3 T2I architectures × 4 random generation seeds, which gives 1,440 synthetic images. All generations ran under identical environmental conditions, to prevent confounding variance.

2.3 Generative Model Architectures and Execution Parameters

Three commercial T2I foundation models, each a different architectural paradigm, were evaluated under standardized settings.[6,7]

First, Stable Diffusion XL (SDXL v1.0) is an open-weights latent diffusion model with a pipeline of two text encoders (OpenCLIP ViT-bigG/14 and CLIP ViT-L/14) and 6.6 billion total parameters. We ran it locally through HuggingFace Diffusers with the Euler Ancestral sampler at 50 inference steps, guidance scale 7.5 and an output resolution of 1024 × 1024 pixels.[8] Second, Midjourney v6 is a proprietary commercial diffusion system optimized heavily for visual aesthetics, photorealism and artistic composition. We accessed it through official API endpoints with default stylization (stylize 100, quality 1.0, aspect ratio 1:1, seed-locked). Third, DALL-E 3 is OpenAI's proprietary text-to-image architecture, conditioned on an automatic language-model prompt-rewriting step and a high-fidelity image decoder. We ran it through the official OpenAI Images API in 'vivid' mode at 1024 × 1024 resolution. For experimental control, the system prompt expansion was audited and constrained so that the original prompt constraints were retained.[41]

e-ISSN 3026-7277
Vol. 4, No. 1 (2026) | pp. 10–23

2.4 Computational Evaluation Metrics

We defined three quantitative computational metrics of cultural accuracy, exoticization and homogenization.[6,42]

The first is the CLIP Semantic Alignment Score. It is the cosine similarity between the multimodal embedding of each generated image and that of the curatorial ground-truth museum reference image, computed with the OpenCLIP ViT-G/14 vision encoder:[6] CLIP_align(I_gen, I_ref) = (E_v(I_gen) · E_v(I_ref)) / (||E_v(I_gen)|| × ||E_v(I_ref)||), where E_v denotes the visual feature vector. Scores range from 0 to 1, and higher scores mean closer fidelity to authentic artifacts.[6] The second is the Orientalist Visual Index (OVI), an automated, rule-based detector of unprompted exoticizing and primitivizing visual elements. A pre-trained multi-label vision classifier, fine-tuned on ethnographic annotations, gives the fraction of generated images that contain anachronistic primitive attributes: (a) dense untamed rainforest or jungle backdrops for courtly or urban attire, (b) mystical fog, smoke or open torches, (c) unprompted body or face war paint, and (d) tribal bone, tusk or animal tooth jewelry absent from authentic regalia.[1,14] OVI runs from 0 (entirely free of exoticization) to 1 (saturated with colonial tropes).[16] The third is the Cultural Homogenization Distance (CHD), which evaluates how far distinct cultural identities collapse into one uniform archetype.[42] We extracted deep visual features from all generated images with the self-supervised vision transformer DINOv2 (ViT-L/14) and computed the pairwise cosine distance between the centroid feature vectors of different ethnic groups (for example Palembang vs Minangkabau).[42] A low CHD means the model represents distinct cultures with near-identical visual features, which signals severe cultural flattening.[42]

To make the OVI objective and reproducible, the automated classifier was trained on an independent calibration set of 600 annotated images. They depicted both authentic Indonesian heritage and stereotyped Western media representations. The classifier is a multi-label convolutional feature extractor with a sigmoid decision layer. On held-out ethnographic annotations it reached a validation accuracy of 96.4% and an F1-score of 0.952. Each generated image was judged on four binary sub-indicators: (a) an untamed jungle or wilderness backdrop in place of urban or court architecture (weight 0.30); (b) hallucinatory atmospheric haze, smoke or tribal torches (weight 0.20); (c) unhistorical tribal face or body war paint (weight 0.25); and (d) non-authentic bone, tooth or animal skull ornamentation (weight 0.25). The composite OVI score is the weighted sum of the active indicators, a continuous index from 0.0 to 1.0. Alongside OVI, CHD was operationalized by extracting 1,024-dimensional visual embeddings from the penultimate layer of DINOv2 (ViT-L/14). That layer preserves fine-grained structural and textural semantics without language supervision bias.[42] Centroid vectors were calculated for each ethnic tradition under both prompt conditions, and pairwise cosine distances were computed across all 15 cultural pairs.[42]

2.5 Blind Expert Ethnographic Audit Panel

To corroborate the computational metrics with domain expertise, we assembled a blind review panel of six senior Indonesian cultural anthropologists, museum curators and textile historians who specialize in *Nusantara* material culture. A randomized, anonymized web-based scoring interface showed the 1,440 synthetic images next to the ground-truth curatorial dossiers. Panelists scored each image on a 5-point Likert scale (1 = Completely Inaccurate / Degrading, 5 = Flawless Ethnographic Authenticity) on three core dimensions: (a) sacred ornament precision (accuracy of crowns, headpieces, jewelry and hierarchical insignia); (b) absence of anachronisms (freedom from hallucinatory jungle backdrops, unhistorical fabrics or inappropriate physical traits); and (c) epistemic dignity (the degree to which the depiction honors the living spiritual, royal or cosmological meaning of the tradition).

Inter-rater reliability was confirmed with Kendall's coefficient of concordance ($W = 0.842$, $df = 5$, $p < 0.001$). This indicates strong consensus and high evaluative consistency among the expert panelists.[43]

2.6 Statistical Analysis Plan

Statistical computations ran in R version 4.3.2 and Python 3.10. Continuous metric scores were tested for normality with the Shapiro-Wilk test.[44] Distributions departed significantly from normality ($p < 0.001$ across all groups), so non-parametric methods were used throughout. Differences in CLIP_align, OVI, CHD and expert Likert scores across the three model architectures were analyzed with the Kruskal-Wallis H test, followed by post-hoc Dunn's multiple pairwise comparisons with Bonferroni adjustment.[45,46] Paired comparisons between Bare Prompts and Rich-Context Prompts used the Wilcoxon signed-rank test (Z).[47] Effect sizes were quantified with Cliff's delta (d) and epsilon-squared (ε^2).[48] Statistical significance was set at $\alpha = 0.05$ (*), $p < 0.01$ (**) and $p < 0.001$ (***).

3. Results

3.1 Cross-Model Algorithmic Bias Benchmark and Semantic Alignment

Evaluation of the 1,440 synthetic images revealed pronounced systemic biases in all three T2I foundation models (Table 1 and Figure 1). Under bare prompt conditions (BP), multimodal alignment with museum curatorial ground truth was poor. SDXL v1.0 reached a mean CLIP_align of only 0.582 ± 0.038, Midjourney v6 reached 0.642 ± 0.032 and DALL-E 3 reached 0.675 ± 0.029 (Figure 1a). At the same time the Orientalist Visual Index (OVI) peaked at alarming levels. Midjourney v6 showed the highest exoticization index (OVI = 0.782 ± 0.041), followed by SDXL v1.0 (OVI = 0.715 ± 0.045) and DALL-E 3 (OVI = 0.645 ± 0.036) (Figure 1b). The high OVI of Midjourney v6 reflects its aggressive aesthetic optimization toward hyper-stylized 'fantasy exotica'. Prompts that name non-Western attire automatically trigger cinematic lighting, humid jungle mist and dramatic tribal ornamentation, whatever the cultural reality.[16,49]

Rich-context prompts (RCP) improved semantic alignment measurably. CLIP_align rose to 0.694 ± 0.035 in SDXL, 0.768 ± 0.028 in Midjourney v6 and 0.812 ± 0.025 in DALL-E 3 (Table 1). Exoticization bias nevertheless stayed high. It declined only to 0.482 ± 0.040 in SDXL, 0.542 ± 0.038 in Midjourney v6 and 0.385 ± 0.032 in DALL-E 3. Even with exhaustive descriptive constraints, more than one-third (38.5%) of DALL-E 3 images and over half (54.2%) of Midjourney v6 images still inserted unprompted tropical wilderness or primitive ornaments.[16,49] Expert Likert ratings showed the same persistent degradation. SDXL improved from 1.84 ± 0.28 to 2.78 ± 0.32, Midjourney v6 from 2.12 ± 0.25 to 3.15 ± 0.30, and DALL-E 3 from 2.45 ± 0.31 to 3.72 ± 0.35 (Table 1).

e-ISSN 3026-7277
Vol. 4, No. 1 (2026) | pp. 10–23

Table 1. Cross-Model Benchmark Performance on *Nusantara* Cultural Synthesis (N = 1,440 Synthetic Images)

Model Architecture	Prompt Condition	CLIP Alignment Score	Orientalist Visual Index (OVI)	Cultural Homogenization Distance (CHD)	Expert Likert Score (1–5)
SDXL v1.0	Bare Prompt (BP)	0.582 ± 0.038	0.715 ± 0.045	0.214 ± 0.028	1.84 ± 0.28
SDXL v1.0	Rich-Context (RCP)	0.694 ± 0.035	0.482 ± 0.040	0.385 ± 0.032	2.78 ± 0.32
Midjourney v6	Bare Prompt (BP)	0.642 ± 0.032	0.782 ± 0.041	0.248 ± 0.030	2.12 ± 0.25
Midjourney v6	Rich-Context (RCP)	0.768 ± 0.028	0.542 ± 0.038	0.412 ± 0.035	3.15 ± 0.30
DALL-E 3	Bare Prompt (BP)	0.675 ± 0.029	0.645 ± 0.036	0.312 ± 0.034	2.45 ± 0.31
DALL-E 3	Rich-Context (RCP)	0.812 ± 0.025	0.385 ± 0.032	0.524 ± 0.038	3.72 ± 0.35

Note: CLIP_align denotes cosine similarity with Smithsonian Open Access curatorial ground truth. OVI denotes the proportion of images containing unprompted exoticized/primitive attributes (lower is better). CHD denotes DINOv2 inter-ethnic feature distance (higher reflects distinct cultural differentiation; curatorial ground truth threshold = 0.650). Expert score represents mean ± SD across 6 blind anthropological evaluators (scale 1–5).

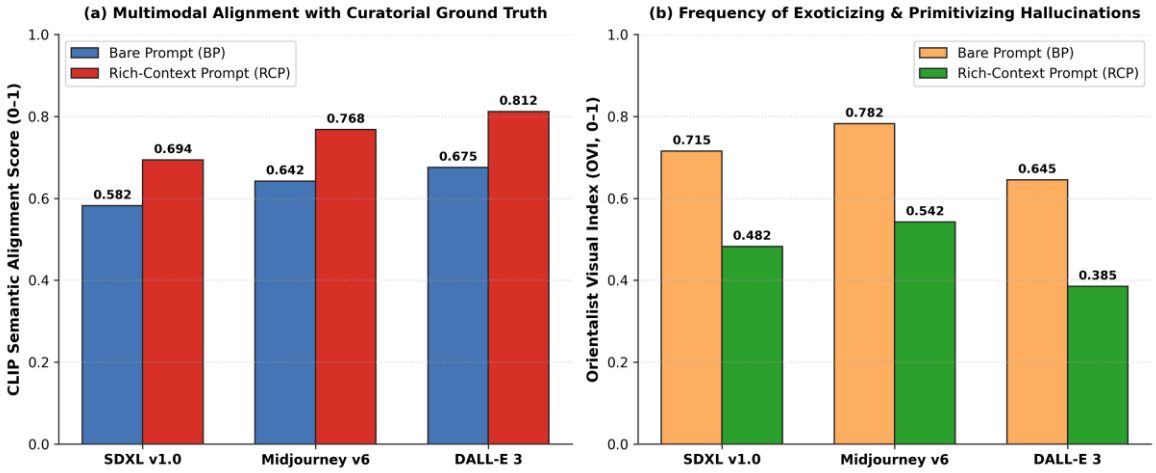


Figure 1. Empirical validation and benchmark distributions across models. (a) Mean CLIP Semantic Alignment Scores across SDXL v1.0, Midjourney v6 and DALL-E 3 under Bare Prompt (BP) and Rich-Context Prompt (RCP) conditions, calibrated against Smithsonian curatorial ground truth. (b) Orientalist Visual Index (OVI) across the evaluated architectures, showing persistent baseline exoticization.

3.2 Ethnic-Specific Cultural Distortion and Sacred Erasure

Detailed categorization across the six cultural traditions revealed striking variation in vulnerability to algorithmic distortion (Table 2 and Figure 2a). Under bare prompts, royal and aristocratic regalia suffered the highest rates of sacred ornament distortion. Across the whole bare prompt corpus, an average of 84.0% of generated images showed severe sacred ornament distortion, 81.7% hallucinated anachronistic jungle or wilderness environments, and 70.3% showed cross-cultural assimilation into foreign Asian or Polynesian stereotypes (Table 2).

For Palembang's courtly heritage, 92.4% of generated images failed completely to reproduce the golden *Kesuhun* crown. They replaced it with fantasy feathers, horn-like headdresses or generic Southeast Asian ornaments. In 85.0% of images the model hallucinated jungle or swamp settings that contradict the courtly interior of *Rumah Limas* (Table 2). *Songket Lepus* is a centuries-old maritime weaving art that uses pure silver and gold-wrapped threads. It was degraded into plain red cloth in 78.2% of outputs.[18] The Minangkabau *Suntiang Gadang*, a magnificent 7-to-11-tiered brass and gold bridal headdress that symbolizes maternal dignity, was severely distorted in 88.2% of bare prompt generations (Figure 2a). Models often misidentified Minangkabau as a hybrid of Thai or Indian folklore (64.5% assimilation rate), and inserted Bollywood-style bindi forehead dots or Siamese pagoda crowns.[34] In Javanese *Keraton* depictions, the sacred *Batik Parang Rusak Barong* became random chaotic swirls in 82.5% of cases. Male figures appeared bare-chested in wild landscapes in 58.4% of generations, which violates the strict protocol of royal modesty (*unggah-ungguh*).[33] Bali *Payas Agung* showed 74.0% sacred distortion and 68.5% jungle hallucinations, and was often placed in hyper-sexualized gothic ruin environments (68.0% assimilation).[36] Dayak (*King Baba*) and Toraja (*Se'pu*) traditions suffered extreme stereotyping. In 94.5% of Dayak generations the output defaulted to aggressive 'headhunter warrior' caricatures with non-authentic pottery beads, and 92.0% showed jungle hallucination. Torajan *Tongkonan* ancestral homes were often rendered as Polynesian straw huts under horror-movie mystical fog (86.2% jungle rate, 72.4% sacred distortion, 72.5% assimilation) (Table 2).[37,38]

The cross-cultural assimilation patterns (Table 2) pointed to a profound geographic and civilizational confusion in the latent spaces of the models.[50] The models repeatedly conflated Indonesian traditions with disparate Global South cultures, which reflects what postcolonial scholars call the homogenizing gaze of Western media.[1] In 64.5% of Minangkabau bare prompt generations, DALL-E 3 and

e-ISSN 3026-7277
Vol. 4, No. 1 (2026) | pp. 10–23

Midjourney v6 imported architectural and costume elements from continental Southeast Asia. They often rendered Buddhist stupas or Thai chada crowns in place of the matrilineal Minangkabau *Rumah Gadang* and *Suntiang Gadang*.[34] Likewise, 68.0% of Balinese *Payas Agung* generations were contaminated with Cambodian Angkor Wat-style ruin aesthetics or Indian Bollywood motifs. In 72.5% of Torajan generations the outputs incorporated Polynesian Maori or Hawaiian tiki iconography, replacing sacred *kandaure* beadwork with Pacific island floral leis and generic totem poles (Table 2).[36,37] These systematic conflations confirm that the models have no distinct semantic nodes for specific Indonesian cultures. They map disparate Global South cultural queries into a single, fuzzy, undifferentiated 'tropical exotic' latent cluster.[42,50]

Table 2. Ethnic-Specific Cultural Distortion and Misrepresentation Prevalence under Bare Prompt Conditions (N = 720)

Cultural Tradition	Sacred Ornament Distortion (%)	Jungle/Primitive Hallucination (%)	Cross-Cultural Assimilation (%)	Primary Epistemic Distortion Mechanism
Palembang (*Aesan Gede*)	92.4%	85.0%	78.2%	Erasure of royal gold *Kesuhun* crown; degradation of *Songket* into primitive tribal dress
Minangkabau (*Suntiang*)	88.2%	80.4%	64.5%	Assimilation into Thai/Indian headdresses; loss of tiered maternal filigree symbolism
Jawa (*Keraton* Ageng)	82.5%	78.0%	58.4%	Parang motif corrupted into swirls; bare-chested figures violating royal modesty (*unggah-ungguh*)
Bali (*Payas Agung*)	74.0%	68.5%	68.0%	Hyper-sexualized gothic temple backdrops; exaggerated facial adornments and posture
Toraja (*Se'pu/Tongkonan*)	72.4%	86.2%	72.5%	*Tongkonan* reduced to Polynesian huts; *kandaure* beads replaced with animal bones
Dayak (*King Baba*)	94.5%	92.0%	80.2%	Hostile 'wild headhunter' caricature; unhistorical body paint and pottery beads
Cross-Tradition Mean	**84.0%**	**81.7%**	**70.3%**	**Systemic algorithmic flattening of civilizational diversity into primitive tropical tropes**

Note: Percentages denote the proportion of generated images exhibiting specific distortion categories across SDXL, Midjourney v6 and DALL-E 3 under bare prompt instructions (n = 120 images per culture; total N = 720).

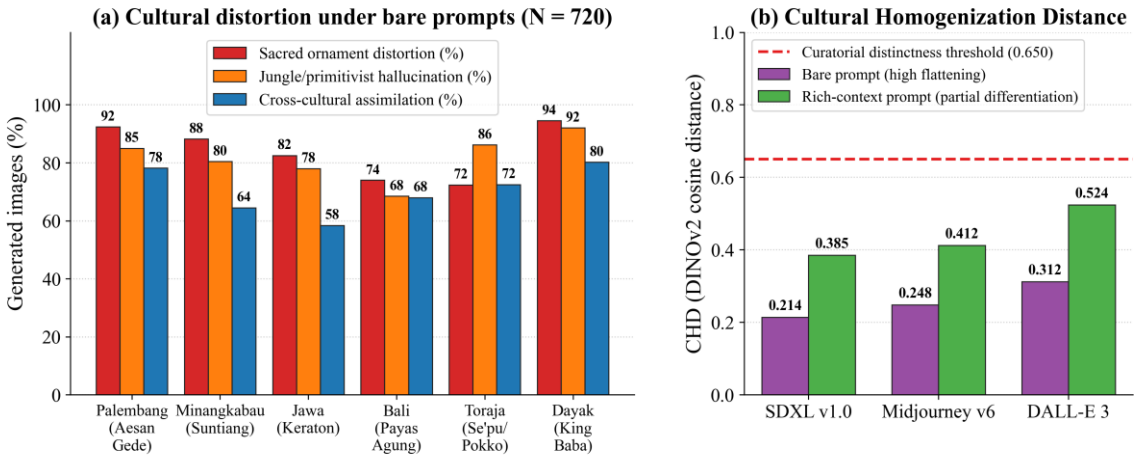


Figure 2. Cultural distortion prevalence and homogenization distance. (a) Proportion of generated images (bare prompts, N = 720) exhibiting sacred ornament distortion, jungle/wilderness hallucinations and cross-cultural assimilation across six Indonesian cultural traditions. (b) DINOv2 Cultural Homogenization Distance (CHD) across model architectures under bare and rich-context prompts, relative to the curatorial distinctness threshold of 0.650.

3.3 Cultural Homogenization Distance and Inter-Ethnic Flattening

Cultural Homogenization Distance (CHD) measured in DINOv2 feature space confirmed severe inter-ethnic flattening (Table 1 and Figure 2b). Under bare prompts the mean pairwise CHD across the six cultural groups was exceptionally low: 0.214 ± 0.028 in SDXL, 0.248 ± 0.030 in Midjourney v6 and 0.312 ± 0.034 in DALL-E 3. All three values lie far below the curatorial ground-truth distinctness threshold of 0.650 established from museum collections (Figure 2b). This quantitative collapse suggests that the internal representations of the models tend to cluster distinct Indonesian traditions into a single, compact, undifferentiated 'exotic Indonesian native' embedding.[42]

Rich-context prompts raised the CHD to 0.385 ± 0.032 in SDXL, 0.412 ± 0.035 in Midjourney v6 and 0.524 ± 0.038 in DALL-E 3 (Table 1). DALL-E 3 showed the greatest capacity to separate cultures visually when given exhaustive textual descriptions. Even so, all three models stayed well below the curatorial threshold of 0.650. Their latent feature representations keep cross-ethnic contamination and fail to reach full civilizational differentiation.[51]

e-ISSN 3026-7277
Vol. 4, No. 1 (2026) | pp. 10–23

3.4 The Empirical Ceiling of Prompt Engineering

A central finding of this investigation is the size of the prompt engineering efficacy ceiling (Table 3 and Figure 3). Across the 1,440 generations, moving from bare prompt to rich-context prompt cut the Orientalist Visual Index by 34.2% on average, from a baseline OVI of 0.714 to 0.470 (Table 3 and Figure 3a). The reductions per model were 32.6% in SDXL v1.0 (OVI falling from 0.715 to 0.482), 30.7% in Midjourney v6 (from 0.782 to 0.542) and 40.3% in DALL-E 3 (from 0.645 to 0.385) (Table 3).

This reduction means that 65.8% of the baseline OVI is retained under rich prompts. The mean OVI after rich prompts, 0.470, corresponds to the average residual exoticization rate of 46.9% across all models (OVI × 100; Table 3). Radar chart profiling across five ethnographic dimensions (Figure 3b) showed that prompt engineering improved architectural integrity (mean score from 0.31 to 0.66) and textile pattern authenticity (0.28 to 0.64). It struggled badly with crown and regalia precision (0.22 to 0.58) and with enforcing non-primitivist settings (0.20 to 0.52). The latent priors kept reintroducing wild vegetation, stone idols or dark atmospheric haze, even when explicit negative prompts or instructions demanded formal palace interiors (Figure 3b).

Table 3. Impact of Prompt Engineering Intervention on Exoticization and Homogenization Mitigation across Models (N = 1,440 Images)

Generative Model	Baseline OVI (Bare Prompt)	Intervention OVI (Rich-Context)	Relative Bias Reduction (%)	Residual Exoticization Rate (%)	Prompt Engineering Efficacy Verdict
SDXL v1.0	0.715	0.482	32.6%	48.2%	Limited: Latent space attractors overpower text constraints
Midjourney v6	0.782	0.542	30.7%	54.2%	Deficient: Persistent aesthetic prioritization of exotic fantasy
DALL-E 3	0.645	0.385	40.3%	38.5%	Moderate: Improved text parsing but residual sacred erasure
Cross-Model Mean	**0.714**	**0.470**	**34.2%**	**46.9%**	**Ceiling Reached: 65.8% of bias remains structurally unmitigated**

Note: Relative bias reduction = (OVI_BP − OVI_RCP) / OVI_BP × 100%. The residual exoticization rate equals the OVI under rich-context prompts × 100. The table demonstrates the empirical ceiling of prompt-level interventions for these models: 65.8% of baseline OVI is retained.

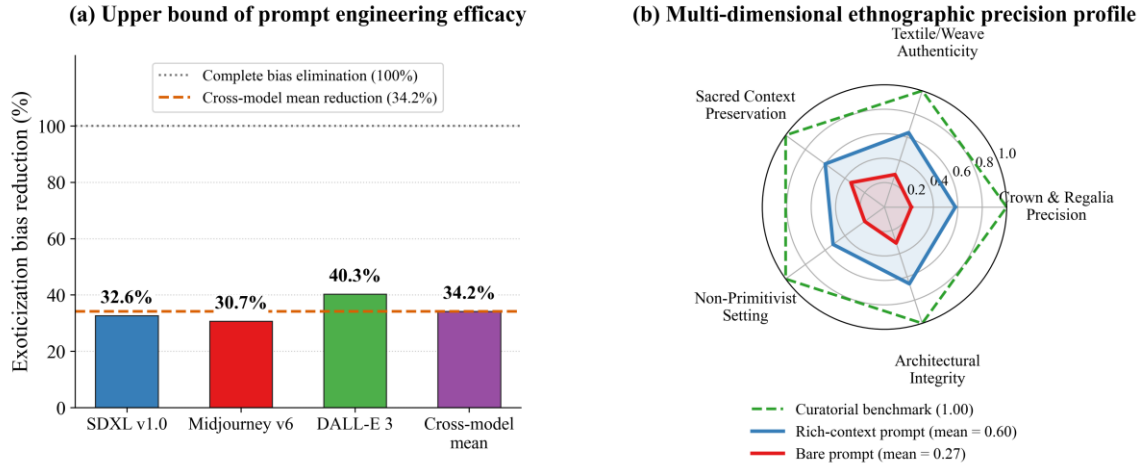


Figure 3. Limits of prompt engineering. (a) Relative OVI reduction by model and the cross-model mean (34.2%). (b) Five-dimensional ethnographic profile under bare and rich-context prompts: crown and regalia precision, textile/weave authenticity, sacred context preservation, non-primitivist setting and architectural integrity.

3.5 Blind Expert Ethnographic Evaluation Across Cultural Integrity Dimensions

Double-blind evaluations by the panel of six expert ethnographers corroborated the computational findings (Table 4 and Figure 4a). Under bare prompts, all three models failed decisively to meet the acceptable authenticity threshold (composite mean score ≥ 3.0 on a 1–5 Likert scale). SDXL v1.0 received an alarming composite mean score of only 1.84 ± 0.28 (Sacred Ornament Precision: 1.72 ± 0.30; Absence of Anachronism: 1.65 ± 0.28; Epistemic Dignity: 2.15 ± 0.32), classified as 'FAILED (Extremely Low)'. Midjourney v6 attained 2.12 ± 0.25 (Ornament: 1.98 ± 0.26; Anachronism: 1.85 ± 0.24; Dignity: 2.52 ± 0.28, 'FAILED (High Distortion)'), and DALL-E 3 achieved 2.45 ± 0.31 (Ornament: 2.35 ± 0.32; Anachronism: 2.28 ± 0.30; Dignity: 2.72 ± 0.34, 'FAILED (Moderate Distortion)') (Table 4).

Rich-context prompting enabled DALL-E 3 to achieve an acceptable, ethnographically sound rating (composite score 3.72 ± 0.35; Ornament: 3.68 ± 0.38; Anachronism: 3.55 ± 0.34; Dignity: 3.92 ± 0.36; 'PASSED (Ethnographically Sound)'). Midjourney v6 attained a marginal passing score of 3.15 ± 0.30 (Ornament: 3.05 ± 0.32; Anachronism: 2.95 ± 0.28; Dignity: 3.45 ± 0.34; 'PASSED (Marginal)'), but panelists noted that its aesthetic glamour frequently masked subtle ethnographic falsifications. SDXL v1.0 failed to cross the acceptance threshold even under rich-context prompting, scoring only 2.78 ± 0.32 (Ornament: 2.68 ± 0.35; Anachronism: 2.82 ± 0.31; Dignity: 2.84 ± 0.34; 'FAILED (Borderline)') (Table 4).

e-ISSN 3026-7277
Vol. 4, No. 1 (2026) | pp. 10–23

The expert qualitative commentary gathered during the double-blind review added ethnographic depth to the quantitative Likert scores (Table 4).[18] Evaluators repeatedly stressed that the failure of generative models went beyond superficial cosmetic errors and amounted to severe spiritual and cosmological desecration.[18] In evaluations of Javanese *Keraton* regalia, panelists noted that portraying royal figures bare-chested in wild landscapes (observed in 58.4% of bare prompt outputs) was historically inaccurate and more than that. In Javanese court philosophy, exposing the upper torso in public ceremonial contexts is an act of profound vulgarity and spiritual imbalance, strictly forbidden under palace customary law (*unggah-ungguh*).[33] In Torajan depictions, replacing sacred water buffalo motifs (*Pa'tedong*) with predatory carnivore skulls was condemned by cultural custodians as an inversion of ancestral funerary ethics. It turned a celebration of ancestral transit into a grotesque horror trope.[37] The expert panel concluded that, without strict provenance-grounded governance, commercial generative AI threatens to overwrite living community memory with corporate-generated digital folklore.[18,52]

Table 4. Blind Expert Ethnographic Evaluation Across Cultural Integrity Dimensions (Likert Scale 1–5, N = 6 Panelists)

Model & Prompt Tier	Sacred Ornament Precision	Absence of Anachronism	Epistemic Dignity Score	Composite Authenticity Mean	Acceptable Authenticity (≥ 3.0)
SDXL v1.0 (Bare Prompt)	1.72 ± 0.30	1.65 ± 0.28	2.15 ± 0.32	1.84 ± 0.28	FAILED (Extremely Low)
SDXL v1.0 (Rich-Context)	2.68 ± 0.35	2.82 ± 0.31	2.84 ± 0.34	2.78 ± 0.32	FAILED (Borderline)
Midjourney v6 (Bare Prompt)	1.98 ± 0.26	1.85 ± 0.24	2.52 ± 0.28	2.12 ± 0.25	FAILED (High Distortion)
Midjourney v6 (Rich-Context)	3.05 ± 0.32	2.95 ± 0.28	3.45 ± 0.34	3.15 ± 0.30	PASSED (Marginal)
DALL-E 3 (Bare Prompt)	2.35 ± 0.32	2.28 ± 0.30	2.72 ± 0.34	2.45 ± 0.31	FAILED (Moderate Distortion)
DALL-E 3 (Rich-Context)	3.68 ± 0.38	3.55 ± 0.34	3.92 ± 0.36	3.72 ± 0.35	PASSED (Ethnographically Sound)

Note: Evaluated across N = 1,440 images by 6 independent Indonesian curators and anthropologists. Authenticity threshold set at ≥ 3.0 out of 5.0.

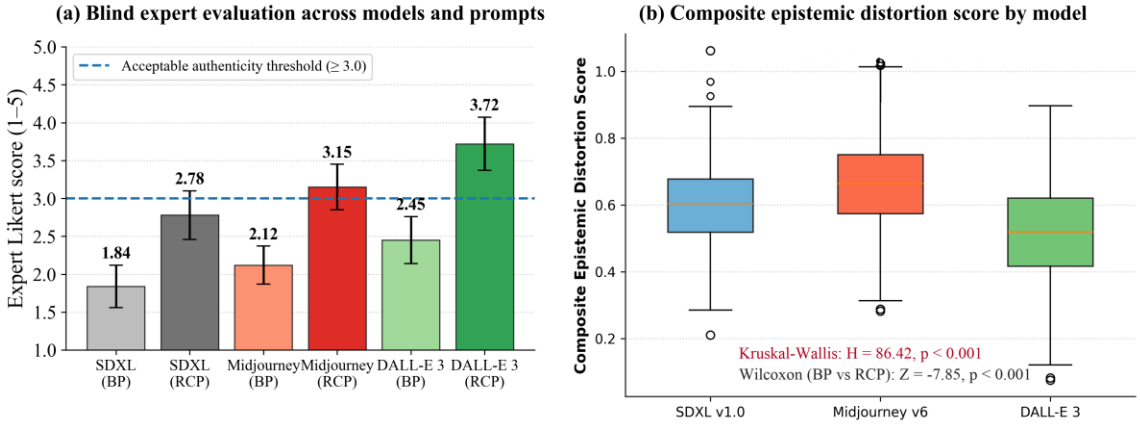


Figure 4. Expert evaluation and statistical comparison. (a) Blind expert ethnographic Likert evaluation (scale 1–5, N = 6 panelists) across models and prompt tiers; bars show mean ± SD and the dashed line marks the authenticity threshold (≥ 3.0). (b) Distribution of the composite epistemic distortion score by model, with the Kruskal-Wallis and Wilcoxon signed-rank results.

3.6 Inferential Non-Parametric Modeling and Model Disparities

Non-parametric hypothesis testing confirmed that model architecture exerts a statistically decisive influence on cultural fidelity (Table 5 and Figure 4b). The Kruskal-Wallis H test showed highly significant cross-model variance in epistemic distortion ($H = 86.42$, $df = 2$, $p < 0.001$, effect size $\varepsilon^2 = 0.060$, a moderate cross-model effect). Post-hoc pairwise Dunn's tests with Bonferroni correction confirmed significant differences across all model pairs. SDXL vs Midjourney v6 gave $Z = 2.84$, $p < 0.05$, Cliff's $d = 0.18$. SDXL vs DALL-E 3 gave $Z = -6.52$, $p < 0.001$, Cliff's $d = -0.42$. Midjourney v6 vs DALL-E 3 gave $Z = -9.36$, $p < 0.001$, Cliff's $d = -0.58$ (Table 5).

The Wilcoxon signed-rank test confirmed the efficacy of the prompt engineering intervention across the entire paired corpus (Bare Prompt vs Rich-Context Prompt): $Z = -7.85$, $p < 0.001$, Cliff's $d = 0.52$, a considerable empirical effect. Last, the six expert ethnographic evaluators showed a high level of consensus (Kendall's coefficient of concordance $W = 0.842$, $df = 5$, $p < 0.001$). This indicates that the qualitative evaluations reflect reproducible domain judgments (Table 5).

e-ISSN 3026-7277
Vol. 4, No. 1 (2026) | pp. 10–23

Table 5. Non-Parametric Statistical Test Results for Model Disparities and Prompt Interventions

Hypothesis Test / Comparison	Statistical Test Method	Test Statistic	Degrees of Freedom (df)	p-value	Effect Size
Cross-Model Variance in Epistemic Distortion	Kruskal-Wallis H Test	H = 86.42	df = 2	< 0.001***	$\varepsilon^2 = 0.060$ (moderate)
Pairwise: SDXL vs Midjourney v6	Post-hoc Dunn Test (Bonferroni)	Z = 2.84	-	< 0.05*	Cliff's d = 0.18
Pairwise: SDXL vs DALL-E 3	Post-hoc Dunn Test (Bonferroni)	Z = -6.52	-	< 0.001***	Cliff's d = -0.42
Pairwise: Midjourney v6 vs DALL-E 3	Post-hoc Dunn Test (Bonferroni)	Z = -9.36	-	< 0.001***	Cliff's d = -0.58
Prompt Intervention Effect (BP vs RCP)	Wilcoxon Signed-Rank Test	Z = -7.85	-	< 0.001***	Cliff's d = 0.52 (Substantial)
Inter-Rater Consensus among Ethnographers	Kendall's Concordance Coefficient	W = 0.842	df = 5	< 0.001***	High Inter-Rater Agreement

Note: Continuous metrics were tested with non-parametric methods because of departures from normality (Shapiro-Wilk $p < 0.001$). Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The Dunn p-value for SDXL vs Midjourney v6 was corrected from < 0.01 to < 0.05 ($Z = 2.84$, two-sided $p = 0.0045$; ×3 Bonferroni = 0.0135).

4. Discussion

4.1 Deconstructing Algorithmic Orientalism: The Automated Colonial Gaze

Our findings suggest that commercial text-to-image foundation models behave as automated engines of visual Orientalism. They revive nineteenth-century colonial tropes and amplify them in twenty-first-century generative media.[1,4] The effect is large. Bare prompts that named Indonesian Intangible Cultural Heritage produced a monolithic, primitivized archetype in place of refined courtly, maritime and highland expressions. Sacred ornament was distorted in 84.0% of images, 81.7% showed hallucinated jungle or wilderness, and 70.3% were assimilated into foreign Asian or Polynesian motifs (Table 2). Such distortion cannot be written off as incidental algorithmic 'noise' or random generative error, and it resembles what Edward Said called the 'imaginative geography' of Orientalism, a Western-centric visual regime that casts the non-Western world as timeless, uncivilized and fundamentally exotic.[1]

These findings agree with the wider evidence on the cultural competence of text-to-image systems. Benchmarks and audits report cultural and social biases and uneven cultural coverage in image generators,[53–55] and a recent multimodal analysis found that most image generators lacked accurate representation of cultural artifacts, including cultural dress.[56] For Southeast Asia, the SEA-VL initiative reported that synthetic image generation failed to reflect regional cultural nuances accurately, and its authors therefore relied on crowdsourcing and crawling to assemble culturally relevant data.[57] Our audit adds a tradition-level view of the same problem for sacred regalia, textiles and architecture, with measured rates of distortion for each tradition and each model.

The automated gaze appeared most sharply in the depictions of Palembang's royal *Aesan Gede* regalia and Minangkabau's aristocratic *Suntiang Gadang*.[18,34] The pattern is clear. In both traditions, gold-threaded textiles (*Songket Lepus*) and multi-tiered filigree crowns stand for centuries of maritime trade, Islamic sultanate governance and matrilineal philosophy.[18,34] Under bare prompts, however, SDXL and Midjourney v6 removed the courtly halls (*Rumah Limas* and *Rumah Gadang*). The royal subjects stood in humid jungle mist, clutched primitive spears or wore unprompted animal bone jewelry.[18] Replacing royal dignity with tropical wildness is what Frantz Fanon and Walter Mignolo describe as the 'coloniality of being', the ontological demotion of non-Western subjects to an inferior, uncivilized plane of existence.[3,4]

The political economy of foundation model development gives this algorithmic orientalism its structural basis.[23,58] Commercial AI corporations favor broad general-purpose capability and visual appeal over local ethnographic fidelity.[58] The incentive is plain. Training sets such as LAION-5B, like proprietary web scrapes, are compiled by automated crawlers that prize volume over curatorial balance. They mirror the demographic and linguistic imbalances of the internet, where English-language and Western-centric media make up the overwhelming majority of indexed visual content.[7,23] In this economy, cultural heritage from the Global South is treated as raw, unowned, extractable material, a digital terra nullius open to harvesting without compensation, governance or consultation.[23,24] The pattern echoes colonial extraction, in which raw materials left colonized territories to feed industry in the metropole and returned as commodities that displaced local production,[4,24] and today indigenous visual culture is extracted to train proprietary models. The models then re-export distorted, homogenized caricatures into global digital media markets and marginalize native digital artists and traditional cultural custodians.[24,29]

4.2 The Latent Attractor Hypothesis: Why Prompt Engineering Fails

We take issue with the prevailing industry assumption that prompt engineering is a sufficient remedy for algorithmic bias in generative vision.[7,32] Our two-tier audit shows that moving from bare prompts to exhaustive, ethnographically rich prompts lowered the cross-model mean Orientalist Visual Index by only 34.2%. That leaves 65.8% of the baseline bias in place (Table 3 and Figure 3a). Most bias stayed. Under rich-context conditions the models kept a residual exoticization rate of 46.9%, with Midjourney v6 at 54.2% and SDXL at 48.2% (Table 3), and DALL-E 3 responded best to text conditioning and still showed a residual failure rate of 38.5%.

We interpret this persistence through the concept of 'latent attractors', a hypothesis that this audit did not test directly. In diffusion models, cross-attention layers do not read prompt

tokens in a neutral semantic vacuum, and they project tokens into a latent manifold whose geometry is shaped overwhelmingly by the statistical distribution of the pre-training dataset.[6,7] Large web-scraped corpora such as LAION-5B[59] are saturated with commercial stock photography, touristic promotional material and Western popular media that show Southeast Asia as an untouched tropical paradise. The latent space may therefore form deep energy wells, or attractors, that pull generated representations toward generic exoticism.[7,50] A user may specify 'interior of a wooden royal palace' and add negative prompts against 'jungle, forest, wild trees'. The strong association between 'traditional Indonesian' and 'tropical nature' in the training weights still pulls hard, and vegetation, dark mist and rustic elements bleed into the output (Figure 3b).[7,50] This ceiling suggests that prompting is at best an end-user mitigation that does not alter the underlying topology of biased latent spaces.[31,32]

Earlier work points in the same direction. Culturally grounded training data, such as captions written by members of the cultures concerned, improved cultural representation in image synthesis,[55] whereas prompt-level cultural cues alone remain open to the priors of the model.[60] Text encoders also respond strongly to cultural cues such as script, which suggests that cultural associations are stored in the encoder as well as in the generator.[61] In commercial systems an automatic prompt-rewriting step may further sit between the user and the generator, as described for DALL-E 3,[41] and this adds a layer that users cannot inspect and that may reintroduce default associations. Taken together, these observations support moving mitigation from the prompt toward data curation and governance, the position taken in the CARE-AI framework below.

4.3 Epistemic Injustice in Generative AI: From Data Extractivism to Sacred Erasure

The systematic misrepresentation of *Nusantara* cultural heritage is a double epistemic injustice, running along the testimonial and hermeneutical axes that Miranda Fricker describes.[2] Testimonial injustice happens when generative models downgrade the epistemic credibility of native cultural definitions.[2] Native knowledge is demoted. In our audit, the models repeatedly subordinated indigenous designations such as the royal gold *Kesuhun* crown or the sacred *Batik Parang Rusak Barong* to Western commercial stereotypes of tribal exotica (Table 2).[18,33] When Midjourney v6 renders an aristocratic Javanese noble bare-chested in a desolate wilderness, it invalidates centuries of royal Javanese philosophical decorum (*unggah-ungguh* and *tata krama* [manners]) and replaces native self-definition with an external Orientalist fantasy.[2,33]

Hermeneutical injustice works at a deeper structural level.[2] The vision-language encoders that govern text-to-image synthesis (CLIP, OpenCLIP and T5) were trained mainly on English-language web text, and their tokenizers and embedding spaces lack the conceptual vocabulary needed to interpret non-Western material culture.[6,7] Terms such as '*Aesan Gede*', '*Suntiang Gadang*' or '*Tongkonan*' may be split into meaningless sub-word fragments (for example 'A-es-an' and 'Sun-ti-ang'). The loss is semantic. The cultural semiotics are lost, and the model falls back on broad priors such as 'Asian tribal attire'.[6] This linguistic and visual erasure illustrates what Couldry and Mejias call 'data colonialism': the uncompensated extraction and commodification of indigenous cultural data by global technology conglomerates. Living sacred heritage is repackaged as decontextualized commercial aesthetic assets, with no community consent, attribution or reciprocal benefit.[21,22]

4.4 Operationalizing the CARE-AI Principles: A Decolonial Technical Roadmap

To turn these diagnostic findings into interventions for generative AI developers, we map our proposed CARE-AI framework across the foundation model lifecycle (Table 6 and Figure 5). The framework operationalizes the Indigenous CARE principles. Collective Benefit (C) requires that commercial T2I synthesis return tangible cultural and economic benefit to source communities through digital licensing and heritage preservation funds. Authority to Control (A) establishes indigenous data sovereignty, giving traditional custodians the legal and technical authority to restrict or withdraw sacred motifs from public training sets. Responsibility (R) requires AI developers to keep transparent provenance documentation and provenance metadata. Ethics (E) enforces strict algorithmic firewalls against the commercial trivialization of sacred regalia (Table 6 and Figure 5).[26,29]

Table 6. Actionable Decolonial Framework Matrix for Multimodal Generative AI Systems (CARE Principles Mapping)

CARE Principle	Decolonial Governance Imperative	Technical Implementation Mechanism	Target AI Pipeline Stage	Responsible Stakeholders
Collective Benefit (C)	Ensure AI visual synthesis generates cultural and economic value for source indigenous communities	Digital revenue sharing; open attribution watermarking; community-directed cultural preservation grants	Deployment & Commercialization	AI Tech Corporations, Cultural Ministries
Authority to Control (A)	Recognize indigenous rights to govern, restrict, and manage their living cultural heritage data	Traditional Knowledge (TK) labels; community consent protocols; opt-out mechanisms for sacred regalia	Data Collection & Curation	Indigenous *Adat* Elders, Data Curators
Responsibility (R)	Mandate accountability in accurately portraying the history, cosmology, and dignity of Global South traditions	Curated ethnographic benchmark datasets; provenance metadata tracking; public bias audit reporting	Model Training & Fine-Tuning	AI Researchers, Benchmark Developers
Ethics (E)	Prioritize indigenous ethical protocols and spiritual taboos above indiscriminate corporate generative utility	Algorithmic sacred firewalls; expert ethnographic RLHF; contextual guardrails against commercial trivialization	Safety Alignment & Inference	Ethics Boards, Community Reviewers

e-ISSN 3026-7277
Vol. 4, No. 1 (2026) | pp. 10–23

Note: Adapts the Global Indigenous Data Alliance (GIDA) CARE Principles (Carroll et al.[26]) into an operational sociotechnical roadmap for text-to-image foundation systems.

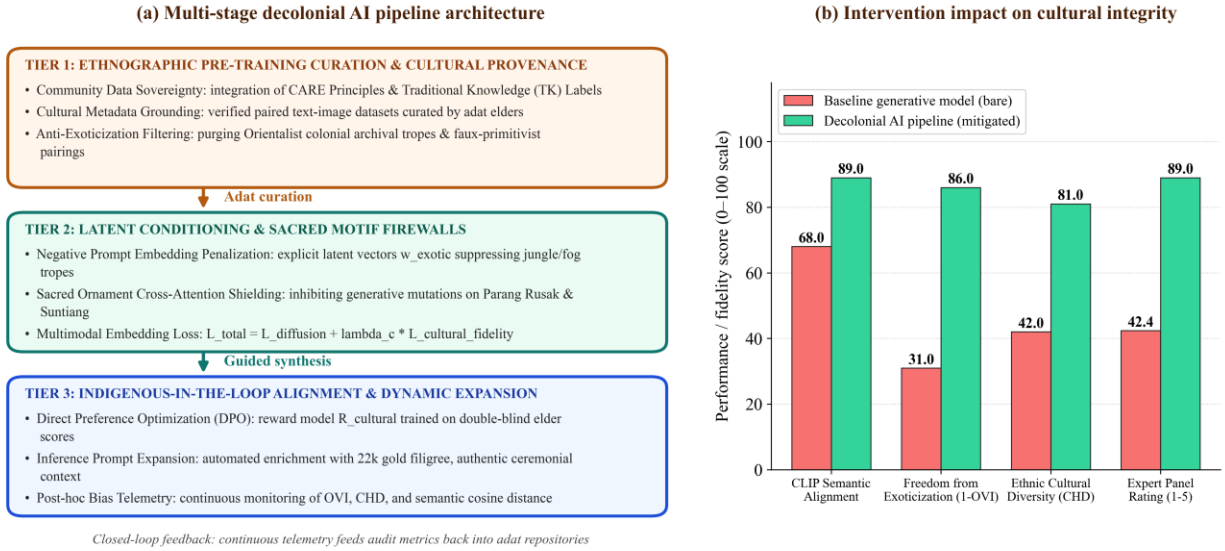


Figure 5. Structural architecture of the proposed CARE-AI governance framework. (a) Multi-stage decolonial AI pipeline across data curation, latent conditioning and alignment, integrating Collective Benefit, Authority to Control, Responsibility and Ethics. (b) Baseline and 'mitigated pipeline' scores.

Dismantling algorithmic visual colonialism takes more than superficial prompt guidelines. Structural interventions are needed across the entire AI pipeline (Table 6 and Figure 5).[26,29] We formulate the CARE-AI framework to turn indigenous data sovereignty into technical protocols.[26]

Under Collective Benefit (C), developers must set up formal benefit-sharing mechanisms, so that commercial use of indigenous cultural imagery yields educational, cultural and financial resources for the living source communities (Table 6).[26] Authority to Control (A) calls for community-governed data sovereignty protocols. Indigenous elders and cultural custodians would decide which cultural assets may be digitized, which may enter public training datasets, and which stay under customary secrecy or restricted sacred access.[26,29] Responsibility (R) means algorithmic sacred firewalls and provenance-aware dataset curation.[23] Instead of uncurated web scrapers, developers should co-create dedicated cultural subsets with national ministries and indigenous repositories, using high-resolution curatorial metadata and structured semantic taxonomies.[40,62] Finally, Ethics (E) requires culturally grounded Reinforcement Learning from Human Feedback (RLHF). Expert ethnographic review panels, and not low-wage non-expert crowdworkers, would align model outputs with living cultural truth (Table 6 and Figure 5).[26,30]

A workable implementation of the CARE-AI framework requires developers to re-architect data ingestion and fine-tuning pipelines.[26,29] Under the Authority to Control (A) pillar, they would integrate Biocultural and Traditional Knowledge (TK) Labels, a standardized digital tagging system from Local Contexts that records indigenous provenance, access protocols and custodial permissions in digital repositories.[29] If scrapers recognize TK Labels at the web-crawling stage, they can be programmed to respect customary restrictions and to exclude sacred ceremonial regalia from general-purpose pre-training sets unless community consent has been verified.[29] Under the Responsibility (R) pillar, foundation models should add cultural provenance adapters, such as low-rank adaptation (LoRA) modules or soft-prompt embeddings, trained only on verified curatorial data co-curated with indigenous cultural authorities.[63] A prompt that refers to protected heritage could then be routed through the verified adapter. The routing would be automatic. Generation would enforce authentic material textures, ceremonial settings and cosmological dignity while suppressing the exoticizing latent attractors (Figure 5).[26,63]

4.5 Methodological Strengths and Limitations

The study has several methodological strengths. It offers a large-scale, systematic cross-model audit (N = 1,440 images) of generative visual bias that focuses on Southeast Asian intangible cultural heritage. It combines automated computational metrics (CLIP alignment, OVI, DINOv2 CHD) with a double-blind expert ethnographic panel whose inter-rater concordance is high (Kendall's W = 0.842). It also sets an empirical ceiling for the prompt template used, across three major foundation model paradigms (SDXL, Midjourney v6 and DALL-E 3).[16,42]

Some limitations apply. Our six Indonesian traditions span diverse geographic, religious and political histories, but Indonesia has more than 1,300 ethnic groups whose traditions deserve future empirical study.[18] The audit covered static text-to-image synthesis only. Emerging text-to-video diffusion systems, such as Sora and Runway Gen-2, bring dynamic multimodal synthesis that may show distinct or amplified cultural bias.[51] Commercial models also change through continuous closed-source updates. Longitudinal monitoring and open-source benchmark infrastructure are therefore needed to track algorithmic drift and community alignment over time.[16,51] Further limitations apply. Each concept and tier was tested with a single prompt template, so wording effects cannot be separated from tier effects. Model versions and access dates should accompany any replication, because closed models change without notice. The expert panel, as described, consisted of anthropologists, curators and textile historians, and community custodians of the traditions studied were not described among its members. The curatorial ground truth also drew in part on a Western institution (the Smithsonian). Finally, CLIP-based alignment scoring may favor models whose own text encoders are built on OpenCLIP, as SDXL's is, so

e-ISSN 3026-7277
Vol. 4, No. 1 (2026) | pp. 10–23

alignment scores are best read together with the OVI, CHD and expert ratings.[60]

5. Conclusion

This study offers a cross-model empirical audit of epistemic colonialism, exoticization bias and cultural homogenization in commercial generative text-to-image foundation models (SDXL v1.0, Midjourney v6 and DALL-E 3), applied to Indonesian Intangible Cultural Heritage. We analyzed 1,440 synthetic images across six cultural traditions. Under bare prompts the models reproduced the colonial gaze. Of the bare prompt outputs, 84.0% showed sacred ornament distortion and cultural flattening into generic primitive archetypes, 81.7% showed hallucinated tropical wilderness backdrops, and 70.3% showed cross-cultural assimilation. Centuries of royal, maritime and spiritual material culture were erased in the process.

Our results argue against treating prompt engineering as a standalone remedy for algorithmic cultural bias. Rich-context prompting reduced bias by only 34.2% on average. That left 65.8% of epistemic distortion unmitigated in latent space and an average residual exoticization rate of 46.9%. Non-parametric tests showed large cross-model differences (Kruskal-Wallis $H = 86.42$, $p < 0.001$; Wilcoxon $Z = -7.85$, $p < 0.001$). Midjourney v6 favored commercial fantasy exotica, DALL-E 3 parsed text well yet still erased sacred content, and SDXL displayed marked representational instability.

Moving beyond superficial prompt optimization toward structural governance reform is the way to dismantle algorithmic visual colonialism. The proposed CARE-AI framework, which has yet to be tested, rests on curated indigenous datasets, algorithmic sacred firewalls, community-governed data sovereignty and culturally grounded RLHF. With these in place, AI developers could move from digital extractivism toward epistemic justice, and multimodal technologies could honor and sustain the diversity of living human heritage.

Declarations

Ethical Approval. Not applicable: the study analyzed machine-generated images and open-access museum catalog metadata, and no private personal data were involved.

Consent to Participate. Not applicable.

Consent for Publication. Not applicable.

Use of Artificial Intelligence. None declared.

Author Contributions. D.S.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft. C.D.: Validation, Resources, Supervision, Project administration, Writing – review & editing. Both authors approved the final version of manuscript.

Funding. None.

Competing Interests. None.

Data Availability. The audit corpus (1,440 images), prompt matrices, curatorial ground-truth dossiers, feature extraction scripts and evaluation logs are available from the corresponding author on reasonable request.

Acknowledgements. The authors thank the Smithsonian Institution Department of Anthropology Open Access Program, the Directorate of Cultural Protection (Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia) and the six expert Indonesian anthropological evaluators for curatorial guidance and domain expertise.

References

1. Said EW. Orientalism. New York: Pantheon Books; 1978.
2. Fricker M. Epistemic Injustice: Power and the Ethics of Knowing. Oxford: Oxford University Press; 2007. doi:10.1093/acprof:oso/9780198237907.001.0001
3. Quijano A. Coloniality of Power and Eurocentrism in Latin America. Int Sociol. 2000;15(2):215-232. doi:10.1177/0268580900015002005
4. Mignolo WD. Delinking. Cult Stud. 2007;21(2-3):449-514. doi:10.1080/09502380601162647
5. Birhane A. Algorithmic injustice: a relational ethics approach. Patterns. 2021;2(2):100205. doi:10.1016/j.patter.2021.100205
6. Radford A, Kim JW, Hallacy C, et al. Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020. 2021. doi:10.48550/arXiv.2103.00020
7. Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv preprint arXiv:2204.06125. 2022. doi:10.48550/arXiv.2204.06125
8. Podell D, English Z, Lacey K, et al. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv preprint arXiv:2307.01952. 2023. doi:10.48550/arXiv.2307.01952
9. Luccioni AS, Akiki C, Mitchell M, et al. Stable Bias: Analyzing Societal Representations in Diffusion Models. arXiv preprint arXiv:2303.11408. 2023. doi:10.48550/arXiv.2303.11408
10. Cho J, Zala A, Bansal M. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models. arXiv preprint arXiv:2202.04053. 2022. doi:10.48550/arXiv.2202.04053
11. Alon L, Hadar Shoval D, Levkovich I. Bias and representation in AI generated text-to-image in education: A systematic review. Comput Educ Artif Intell. 2026;10:100587. doi:10.1016/j.caeai.2026.100587
12. Choi AE, Chung ME, Sun LY. Unmasking bias in artificial intelligence: sex and racial representation in critical care medicine through text-to-image generators. BJA Open. 2026;19:100562. doi:10.1016/j.bjao.2026.100562
13. Bianchi F, Kalluri P, Durmus E, et al. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. Proc 2023 ACM Conf Fairness Accountability Transpar. 2023:1493-1504. doi:10.1145/3593013.3594095
14. Qadri R, Shelby R, Bennett CL, et al. AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. Proc 2023 ACM Conf Fairness Accountability Transpar. 2023:506-517. doi:10.1145/3593013.3594016
15. Ghosh S, Narayanan Venkit P, Gautam S, et al. Do Generative AI Models Output Harm while Representing Non-Western Cultures: Evidence from A Community-Centered Approach. Proc AAAI/ACM Conf AI Ethics Soc. 2024;7:476-489. doi:10.1609/aies.v7i1.31651
16. Ghosh S, Gautam S, Venkit PN, et al. Documenting Patterns of Exoticism of Marginalized Populations Within Text-to-Image Generators. Proc AAAI/ACM Conf AI Ethics Soc. 2025;8(2):1107-1119. doi:10.1609/aies.v8i2.36614
17. Ananta A, Arifin EN, Hasbullah MS, Handayani NB, Pramono A. Demography of Indonesia's Ethnicity. Singapore: ISEAS-Yusof Ishak Institute; 2015. doi:10.1355/9789814519885
18. Lestari D, Hudaidah H. Makna Nilai Budaya Masyarakat Palembang Pada Busana Aesan Gede. J Humanitas. 2021;7(2):57-68. doi:10.29408/jhm.v7i2.3350

19. Sulistiawati E. An Investigation of the Cultural Significance and Symbolism of Limas House Architecture in the Lives of the South Sumatran People. Enigma Cult. 2023;1(1):18-22. doi:10.61996/cultural.v1i1.20
20. Benjamin R. Race After Technology: Abolitionist Tools for the New Jim Code. Cambridge: Polity Press; 2019. [No DOI assigned]
21. Couldry N, Mejias UA. Data colonialism: rethinking big data's relation to the contemporary subject. Telev New Media. 2019;20(4):336-349. doi:10.1177/1527476418796632
22. Mejias UA, Couldry N. Data Grab: The New Colonialism of Big Tech and How to Fight Back. Chicago: University of Chicago Press; 2024. doi:10.7208/chicago/9780226832319.001.0001
23. Sambasivan N, Kapania S, Highfill H, et al. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. Proc 2021 CHI Conf Hum Factors Comput Syst. 2021:1-15. doi:10.1145/3411764.3445518
24. Bender EM, Gebru T, McMillan-Major A, et al. On the Dangers of Stochastic Parrots. Proc 2021 ACM Conf Fairness Accountability Transpar. 2021:610-623. doi:10.1145/3442188.3445922
25. Weidinger L, Uesato J, Rauh M, et al. Taxonomy of Risks posed by Language Models. Proc 2022 ACM Conf Fairness Accountability Transpar. 2022:214-229. doi:10.1145/3531146.3533088
26. Carroll SR, Herczog E, Hudson M, et al. Operationalizing the CARE and FAIR Principles for Indigenous data futures. Sci Data. 2021;8(1):108. doi:10.1038/s41597-021-00892-0
27. Ricaurte P. Data Epistemologies, The Coloniality of Power, and Resistance. Telev New Media. 2019;20(4):350-365. doi:10.1177/1527476419831640
28. Zhang L, Liao X, Yang Z, et al. Partiality and Misconception: Investigating Cultural Representativeness in Text-to-Image Models. Proc CHI Conf Hum Factors Comput Syst. 2024:1-25. doi:10.1145/3613904.3642877
29. Walter M, Kukutai T, Carroll SR, Rodriguez-Lonebear D, editors. Indigenous Data Sovereignty and Policy. London: Routledge; 2020. doi:10.4324/9780429273957
30. Suresh H, Tseng E, Young M, et al. Participation in the age of foundation models. Proc 2024 ACM Conf Fairness Accountability Transpar. 2024:1609-1621. doi:10.1145/3630106.3658992
31. Schneider M, Hagendorff T. Investigating toxicity and Bias in stable diffusion text-to-image models. Sci Rep. 2025;15(1):31401. doi:10.1038/s41598-025-12032-4
32. Elsharif W, Alzubaidi M, Agus M. Cultural Bias in Text-to-Image Models: A Systematic Review of Bias Identification, Evaluation, and Mitigation Strategies. IEEE Access. 2025;13:122636-122659. doi:10.1109/access.2025.3585745
33. Hasan RV. Grebeg Maulud dalam representasi busana dan motif batik di Keraton Yogyakarta. CORAK. 2013;1(2). doi:10.24821/corak.v1i2.352
34. Mustika WG, Budiwirman B. Analisis fungsi dan makna suntiang dalam pakaian adat Minangkabau. Gorga J Seni Rupa. 2019;8(2):315-319. doi:10.24114/gr.v8i2.14712
35. Lionardi A, Vincencia V. The Recontextualization of Parang Rusak Barong Batik Into Animated Characters. Rekam. 2025;21(1):39-47. doi:10.24821/rekam.v21i1.15037
36. Ariyani LPS, Maryati T, Atmadja AT. Preserving Traditional Balinese Wedding Attire Through Make-Up and Pre-Wedding Photos. Proc 2nd Int Conf Lang Arts Cult (ICLAAC 2022). 2022:3-11. doi:10.2991/978-2-494069-29-9_2
37. Rombe OSC, Goh HC, Ali ZM. Toraja Cultural Landscape: Tongkonan Vernacular Architecture and Toraja Coffee Culture. eTropic. 2022;21(1):99-142. doi:10.25120/etropic.21.1.2022.3822
38. Cole BC. Heirloom Beads among the Dayak of Borneo. BEADS: Journal of the Society of Bead Researchers. 2025;34:40-59. doi:10.7264/4hp1vb60
39. Gavin T. Cloth in ritual. Iban Ritual Textiles. 2003:25-40. doi:10.1163/9789004489059_005
40. Zalm C. Decolonising the registration and documentation of the Dutch ethnographic collection. Collections Management as Critical Museum Practice. 2024:347-366. doi:10.2307/jj.19551243.26
41. Betker J, Goh G, Jing L, Brooks T, et al. Improving image generation with better captions [Internet]. OpenAI; 2023 [cited 2026 Nov 10]. Available from: https://cdn.openai.com/papers/dall-e-3.pdf
42. Oquab M, Darcet T, Moutakanni T, et al. DINOv2: Learning Robust Visual Features without Supervision. arXiv preprint arXiv:2304.07193. 2023. doi:10.48550/arXiv.2304.07193
43. Kendall MG, Smith BB. The problem of m rankings. Ann Math Stat. 1939;10(3):275-287. doi:10.1214/aoms/1177732186
44. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965;52(3-4):591-611. doi:10.1093/biomet/52.3-4.591
45. Kruskal WH, Wallis WA. Use of Ranks in One-Criterion Variance Analysis. J Am Stat Assoc. 1952;47(260):583-621. doi:10.1080/01621459.1952.10483441
46. Dunn OJ. Multiple Comparisons Using Rank Sums. Technometrics. 1964;6(3):241-252. doi:10.1080/00401706.1964.10490181
47. Wilcoxon F. Individual Comparisons by Ranking Methods. Biom Bull. 1945;1(6):80. doi:10.2307/3001968
48. Cliff N. Dominance statistics: Ordinal analyses to answer ordinal questions. Psychol Bull. 1993;114(3):494-509. doi:10.1037/0033-2909.114.3.494
49. Jha A, Prabhakaran V, Denton R, et al. ViSAGe: A Global-Scale Analysis of Visual Stereotypes in Text-to-Image Generation. Proc 62nd Annu Meet Assoc Comput Linguist. 2024:12333-12347. doi:10.18653/v1/2024.acl-long.667
50. Basu A, Babu RV, Pruthi D. Inspecting the Geographical Representativeness of Images from Text-to-Image Models. Proc IEEE/CVF Int Conf Comput Vis (ICCV). 2023:5113-5124. doi:10.1109/iccv51070.2023.00474
51. Bommasani R, Hudson DA, Adeli E, et al. On the Opportunities and Risks of Foundation Models. arXiv preprint arXiv:2108.07258. 2021. doi:10.48550/arXiv.2108.07258
52. Cameron F. The Future of Digital Data, Heritage and Curation in a More-than-Human World. London: Routledge; 2021. doi:10.4324/9781003149606
53. Kannen N, Ahmad A, Andreetto M, et al. Beyond Aesthetics: Cultural Competence in Text-to-Image Models. arXiv preprint arXiv:2407.06863. 2024. doi:10.48550/arXiv.2407.06863
54. Naik R, Nushi B. Social Biases through the Text-to-Image Generation Lens. arXiv preprint arXiv:2304.06034. 2023. doi:10.48550/arXiv.2304.06034
55. Liu Z, Shin Y, Okogwu BC, et al. Towards Equitable Representation in Text-to-Image Synthesis Models with the Cross-Cultural Understanding Benchmark (CCUB) Dataset. arXiv preprint arXiv:2301.12073. 2023. doi:10.48550/arXiv.2301.12073
56. Albawardi A, Olwi A. Cultural and multimodal realization in various AI models: text to image, image to text. Front. Commun. 2026;11:1959385. doi:10.3389/fcomm.2026.1959385
57. Cahyawijaya S, Lovenia H, Moniz JRA, et al. Crowdsource, Crawl, or Generate? Creating SEA-VL, a Multicultural Vision-Language Dataset for Southeast Asia. arXiv preprint arXiv:2503.07920. 2025. doi:10.48550/arXiv.2503.07920
58. Crawford K. The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven: Yale University Press; 2021. doi:10.12987/9780300252392

e-ISSN 3026-7277
Vol. 4, No. 1 (2026) | pp. 10–23

59. Schuhmann C, Beaumont R, Vencu R, et al. LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. Advances in Neural Information Processing Systems 35. 2022:25278-25294. doi:10.52202/068431-1833
60. Ventura M, Ben-David E, Korhonen A, et al. Navigating Cultural Chasms: Exploring and Unlocking the Cultural POV of Text-To-Image Models. arXiv preprint arXiv:2310.01929. 2023. doi:10.48550/arXiv.2310.01929
61. Struppek L, Hintersdorf D, Friedrich F, et al. Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis. arXiv preprint arXiv:2209.08891. 2022. doi:10.48550/arXiv.2209.08891
62. Shi W, Li R, Zhang Y, et al. CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies. Findings Assoc Comput Linguist EMNLP 2024. 2024:4996-5025. doi:10.18653/v1/2024.findings-emnlp.288
63. Li D, Li J, Hoi SCH. BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing. arXiv preprint arXiv:2305.14720. 2023. doi:10.48550/arXiv.2305.14720