Statistical Monoculture: How Text-to-Image AI Flattens Cultural and Aesthetic Diversity

CCC researchers; TBA

February 27, 2026

Abstract

Why do AI image generators produce outputs that all look the same? The prevailing explanation blames biased training data, but we show that the problem runs deeper: aesthetic homogenisation is a mathematical consequence of how these models are built. We prove that the standard training objective used in diffusion models—the engine behind DALL-E, Midjourney, Stable Diffusion, and their successors—computes a probability-weighted average over all aesthetic possibilities consistent with a given prompt. Cultural expressions with low statistical mass in the training data are not merely underrepresented; they are *averaged away* by the optimisation itself. We call this the *Mode Averaging Principle* (MAP) and show that it combines with two additional forces—classifier-free guidance, which amplifies averaging in pursuit of "quality," and recursive data contamination, whereby each generation's narrowed outputs enter future training sets—to form a compound convergence system that progressively flattens aesthetic diversity. The observable result is what we term the *Statistical Levelling of Originality Principle* (SLOP): the generic, culturally homogeneous quality increasingly characteristic of AI-generated imagery. We validate these predictions in three complementary ways: synthetic experiments on controlled distributions, empirical analysis of 746 images from OpenAI's DALL-E 3 and Google's Imagen 4, and an ablation study on open-source models that isolates the MSE objective from preference training. The results are striking: commercial models discard over 98% of the aesthetic variation available to them, compressing outputs into roughly 10 effective dimensions out of 768; minority cultural representations are rendered 15% less diversely than majority ones; and two independently developed models—built by different companies on different datasets—converge to 81% similarity for the same prompts, more than 20 standard deviations above chance. The ablation study demonstrates that this homogenisation is not primarily driven by RLHF or preference training: a base model with no preference training already exhibits 91% of the majority/minority diversity gap observed in its preference-trained variant, and classifier-free guidance monotonically amplifies the effect. These findings demonstrate that current text-to-image AI enforces a *statistical monoculture*—not through deliberate design, but through the mathematics of its objective function. The problem is not only in the data; it is in the calculus. We discuss implications for diversity-preserving generative design, identifying intervention points at each of the three convergence forces.

1 Introduction

The rapid proliferation of diffusion-based generative models—from DALL-E [Ramesh et al., 2021] and Stable Diffusion [Rombach et al., 2022] to Midjourney and their successors—has fundamentally transformed the production and consumption of digital visual culture. These systems, built on the foundational work of denoising diffusion probabilistic models [Ho et al., 2020, Sohl-Dickstein et al., 2015] and score-based generative modelling [Song and Ermon, 2019, Song et al., 2021b], can produce images of striking technical proficiency from simple text prompts. Yet alongside their rapid adoption, a persistent critique has emerged from scholars working at the intersection of computer science, cultural studies, and the social sciences: these models

tend to produce outputs with a characteristic, often generic "AI aesthetic"—an aesthetic of smoothness, digital perfection, and stylistic blending that feels simultaneously technically proficient and culturally hollow [Lindgren, 2024, Raley and Rhee, 2023, Roberge and Castelle, 2021, Offert and Dhaliwal, 2025].

This phenomenon, which Roland Meyer (drawing on Jacob Birken) has termed "platform realism" [Meyer, 2025], describes a second-order aesthetic of generic images that are statistically optimised to be legible, plausible, and structurally conservative, often reflecting dominant cultural values. Although generative models are frequently billed as producing "realistic" imagery, in practice they tend toward outputs that are heavily biased towards Western, male, and middle-class aesthetic preferences [Meyer, 2025, Bianchi et al., 2023, Luccioni et al., 2023]. This is attributable not only to the statistical prevalence of such aesthetics in the training data, but also to the filtering processes, RLHF stages, and consumer-oriented design choices embedded in commercial deployment pipelines. Empirical benchmarks such as CultDiff have confirmed that state-of-the-art models frequently fail to generate culturally accurate artefacts for underrepresented regions, pointing to a systematic erasure of non-Western aesthetics [Bayramli et al., 2025].

The critical study of these technologies has developed rapidly under the banner of "Critical AI studies" [Lindgren, 2024, Raley and Rhee, 2023, Roberge and Castelle, 2021], with scholars developing methods for interrogating the social and technical dimensions of generative systems [Offert and Dhaliwal, 2025, Offert and Phan, 2024] and tracing their connections to the actuarial sciences and the politics of ground truth [Amoore, 2023, Kang, 2023, Sadowski, 2025]. However, a gap remains: while qualitative critiques have identified the homogenisation problem with precision, the field lacks a rigorous mathematical account of *why* diffusion models produce homogenised outputs as a structural matter, not merely as a consequence of biased data.

This paper fills that gap. We develop a unified mathematical framework demonstrating that the observed aesthetic homogenisation is driven by three reinforcing convergence forces:

1. **The denoising objective** (the Mode Averaging Principle, MAP): We prove that the standard mean-squared-error training loss compels the model to learn a conditional expectation operator that, under conditions of high noise, converges to a probability-weighted average of aesthetic modes. Modes with greater statistical mass—corresponding to dominant cultural aesthetics—exert disproportionate gravitational pull on this average.

2. **Classifier-free guidance** (CFG): We show that the standard technique used to improve prompt fidelity and perceived "quality" [Ho and Salimans, 2022] mathematically amplifies the mode-averaging effect, narrowing the output distribution around the dominant conditional mean.

3. **Recursive data contamination**: Drawing on recent work on model collapse [Shumailov et al., 2024, Alemohammad et al., 2023], we argue that the already-narrowed outputs of current models are entering the training sets of future models, creating a compound variance collapse across generations.

The observable consequence of this compound convergence system is what we term the *Statistical Levelling of Originality Principle* (SLOP)—the pervasive generic quality, lack of true distinction, and tendency to dilute unique expressions into a bland statistical average that characterises much AI-generated content. We validate these theoretical predictions in three complementary ways: first, through controlled synthetic experiments on mixture distributions where the ground truth is known; second, through empirical analysis of 746 images generated by two leading commercial models—OpenAI's DALL-E 3 [Betker et al., 2023] and Google's Imagen 4 [Google DeepMind, 2025]—using CLIP embeddings [Radford et al., 2021] to measure diversity in semantic space; and third, through an ablation study on open-source models (SDXL) that isolates the contribution of the MSE objective from preference training. The empirical results

confirm all key predictions: within-prompt effective dimension collapses to approximately 10 out of 768 dimensions, minority cultural representations are 15% less diverse than majority ones, and different models converge to cosine similarity 0.81 for the same prompts. The ablation demonstrates that 91% of the majority/minority diversity gap is present in a base model with no preference training, establishing the denoising objective as the primary structural driver. Our research is guided by two core questions:

1. Why do diffusion models inherently produce a specific, "averaged" aesthetic, and what are the precise mathematical mechanisms underlying this homogenisation?
2. How might cultural critiques of AI-generated art proceed from, and be strengthened by, a rigorous technical and mathematical framework?

We proceed as follows. Section 2 establishes the mathematical language for analysing aesthetic distributions. Section 3 contains our central mathematical results, proving the Mode Averaging Principle and extending it to classifier-free guidance and trajectory dynamics. Section 4 develops the compound convergence system that links intra-generation averaging to inter-generation collapse. Section 5 connects our mathematical framework to existing cultural critiques, developing the SLOP concept and its implications. Section 6 presents synthetic experimental validation. Section 7 reports empirical validation on commercial text-to-image models. Section 8 presents an ablation study disentangling the MSE objective from preference training. Section 9 discusses implications, limitations, and future directions.

2 A mathematical framework for aesthetic diversity

To move from qualitative observation to formal analysis, we require a precise mathematical language for describing how generative models shape aesthetic diversity. This section establishes that language, defining distributions over an abstract aesthetic space and introducing measures of their concentration, diversity, and complexity. These measures draw on information theory and statistical mechanics, and will be applied directly to the diffusion process in subsequent sections.

2.1 Aesthetic space and distributions

We conceptualise an *aesthetic space* ($\mathcal{X}$) as a high-dimensional manifold, possibly embedded in $\mathbb{R}^d$, where each point $x \in \mathcal{X}$ represents a distinct aesthetic artefact, style, or cultural representation. A *probability distribution* P over this space, with density $p(x)$, describes the landscape of training data available to a generative model. The density $p(x)$ indicates the relative likelihood of finding aesthetics similar to x in the training corpus. Real-world datasets produce highly non-uniform distributions: certain regions of $\mathcal{X}$ corresponding to dominant cultural modes have much higher probability density than others.

2.2 Distributional measures

Definition 2.1 (Measures for Aesthetic Analysis)**.** *For a probability distribution P on the aesthetic space $\mathcal{X}$ with density $p(x)$, and a parameter $\alpha \in (0, 1)$, we define:*

*α-**Level Set**: The region of highest density containing probability mass α:*

$$L_\alpha(P) = \{x \in \mathcal{X} : p(x) \geq q_\alpha(P)\}$$

where $q_\alpha(P)$ is the $(1-\alpha)$-quantile of p, satisfying $\int_{p(x)\geq q_\alpha(P)} p(x)\,dx = \alpha$.

***Effective Support Radius**: A measure of distributional spread:*

$$R_\alpha(P) = \inf\{r > 0 : \mathbb{P}_{X\sim P}(\|X - \mu_P\|_2 \leq r) \geq \alpha\}$$

where $\mu_P = \mathbb{E}_{X\sim P}[X]$ *is the mean of* P*.*

Diversity Index*: An information-theoretic measure based on differential entropy:*

$$D(P) = \exp(H(P)), \quad \textit{where } H(P) = -\int_{\mathcal{X}} p(x)\log p(x)\,dx.$$

This is the perplexity *of the distribution, representing the effective volume of the distribution's support [Shannon, 1948].*

Effective Dimension*: A measure of distributional complexity:*

$$d_{\text{eff}}(P) = \frac{1}{\int_{\mathcal{X}} p(x)^2\,dx}$$

This is equivalent to $\exp(H_2(P))$ *where* H_2 *is the Rényi entropy of order 2, and is known in statistical physics as the* participation ratio*—the effective number of states contributing to the distribution.*

These definitions use quantiles and information-theoretic measures that are well-defined for general continuous distributions, ensuring our framework handles the high-dimensional aesthetic spaces in which generative models operate.

Definition 2.2 (Measuring Concentration and Homogenisation)**.** *Given an original distribution* P *and a transformed distribution* Q*, we define:*

Support Concentration Ratio*:* $\rho_\alpha(Q,P) = R_\alpha(Q)/R_\alpha(P)$

Diversity Loss Ratio*:* $\delta(Q,P) = D(Q)/D(P)$

Effective Dimension Ratio*:* $\gamma(Q,P) = d_{\text{eff}}(Q)/d_{\text{eff}}(P)$

We say Q *is* ***concentrated relative to*** P *if* $\rho_\alpha(Q,P) < 1$*, exhibits* ***diversity loss*** *if* $\delta(Q,P) < 1$*, and shows* ***dimensional collapse*** *if* $\gamma(Q,P) < 1$*.*

Values below 1 for any of these ratios constitute direct, quantifiable evidence of homogenisation. In what follows, we demonstrate that the diffusion process systematically drives all three ratios below 1.

3 The Mode Averaging Principle: statistical gravity in diffusion models

This section contains the central mathematical results of the paper. We demonstrate that the denoising diffusion objective is not a neutral learning procedure but a mathematical engine for statistical concentration, and we extend this analysis to classifier-free guidance and the dynamics of the sampling trajectory.

3.1 The diffusion process and its objective

A diffusion model [Ho et al., 2020, Song et al., 2021b] operates in two phases. In the *forward process*, Gaussian noise is progressively added to training data x_0 over T timesteps:

$$x_t = \sqrt{\bar{\alpha}_t}\,x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is the cumulative noise schedule, and $\sigma_t^2 = 1 - \bar{\alpha}_t$ is the noise variance at step t.

In the *reverse process*, a neural network $\epsilon_\theta(x_t, t)$ is trained to predict the added noise. The standard training objective is the mean squared error loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0\sim p_{\text{data}},\,\epsilon\sim\mathcal{N}(0,I),\,t}\left[\|\epsilon - \epsilon_\theta(x_t,t)\|^2\right]$$

The optimal predictor minimising this loss is the conditional expectation $\epsilon^*_\theta(x_t, t) = \mathbb{E}[\epsilon \mid x_t]$. By the reparameterisation identity, this is equivalent to the model learning to predict $\mathbb{E}[x_0 \mid x_t]$—the *statistical average* of all original data points that could have produced the observed noisy input x_t.

This fact—that the optimal denoiser computes a conditional expectation—is well known in the machine learning literature. What has not been formalised is its cultural consequence: conditional expectation is, by mathematical construction, an *averaging operator*. It does not select the most likely mode, preserve rare modes, or maintain distributional diversity. It computes a weighted mean. The remainder of this section makes this consequence precise.

3.2 Formalising the training data landscape

Assumption 3.1 (Mixture Model of Aesthetic Data)**.** *The training distribution P_{data} can be approximated as a mixture of k aesthetic modes:*

$$P_{\text{data}}(x) = \sum_{i=1}^{k} \pi_i \, \phi_i(x)$$

where:

- $\pi_i > 0$ *are mixing weights with* $\sum_{i=1}^{k} \pi_i = 1$*, representing the* statistical mass *of each mode in the training data;*
- $\phi_i(x) = \mathcal{N}(x;\, \mu_i,\, \Sigma_i)$ *are Gaussian component densities with means μ_i (the aesthetic centre of each mode) and covariances Σ_i (the internal diversity within each mode);*
- *the components are well-separated: $\|\mu_i - \mu_j\| \geq \delta > 0$ for $i \neq j$.*

This assumption captures the structure of real aesthetic data, which clusters into identifiable styles—"contemporary Western portraiture", "East Asian ink wash painting", "West African textile patterns"—with vastly different prevalences in web-scraped training corpora. The weights π_i encode the power asymmetry: dominant modes (e.g., Western stock photography) have large π_i, while minority modes (e.g., authentic Bangladeshi cultural representation) have small π_i.

3.3 The central theorem: mode averaging under high noise

We first establish a key lemma about the within-component conditional expectation, which the existing literature often leaves implicit.

Lemma 3.2 (Within-Component Shrinkage)**.** *Under the diffusion process $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$, for a Gaussian component C_i with mean μ_i and covariance Σ_i, the within-component conditional expectation is:*

$$\mathbb{E}[x_0 \mid x_t,\, C_i] = \mu_i + \frac{\bar{\alpha}_t \Sigma_i}{\bar{\alpha}_t \Sigma_i + \sigma_t^2 I} \left(\frac{x_t}{\sqrt{\bar{\alpha}_t}} - \mu_i \right)$$

In the high-noise limit ($\sigma_t^2 \to \infty$), the shrinkage factor $\bar{\alpha}_t \Sigma_i / (\bar{\alpha}_t \Sigma_i + \sigma_t^2 I) \to 0$, and therefore:

$$\mathbb{E}[x_0 \mid x_t,\, C_i] \to \mu_i$$

That is, when noise is high, the observation x_t carries vanishing information about the original data point's position within mode C_i, and the conditional expectation collapses to the mode centre.

Proof. Given $x_0 \sim \mathcal{N}(\mu_i, \Sigma_i)$ and $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, the joint distribution of (x_0, x_t) conditional on C_i is Gaussian. The conditional distribution $x_0 \mid x_t, C_i$ is Gaussian with mean given by the standard formula for Gaussian conditioning:

$$\mathbb{E}[x_0 \mid x_t, C_i] = \mu_i + \mathrm{Cov}(x_0, x_t)\, \mathrm{Var}(x_t)^{-1}\, (x_t - \mathbb{E}[x_t \mid C_i])$$

We have $\mathrm{Cov}(x_0, x_t) = \sqrt{\bar{\alpha}_t}\, \Sigma_i$, $\mathrm{Var}(x_t \mid C_i) = \bar{\alpha}_t \Sigma_i + \sigma_t^2 I$, and $\mathbb{E}[x_t \mid C_i] = \sqrt{\bar{\alpha}_t}\, \mu_i$. Substituting:

$$\begin{aligned}\mathbb{E}[x_0 \mid x_t, C_i] &= \mu_i + \sqrt{\bar{\alpha}_t}\, \Sigma_i \left(\bar{\alpha}_t \Sigma_i + \sigma_t^2 I\right)^{-1} \left(x_t - \sqrt{\bar{\alpha}_t}\, \mu_i\right) \\ &= \mu_i + \frac{\bar{\alpha}_t \Sigma_i}{\bar{\alpha}_t \Sigma_i + \sigma_t^2 I} \left(\frac{x_t}{\sqrt{\bar{\alpha}_t}} - \mu_i\right)\end{aligned}$$

As $\sigma_t^2 \to \infty$, the matrix $\bar{\alpha}_t \Sigma_i (\bar{\alpha}_t \Sigma_i + \sigma_t^2 I)^{-1} \to 0$ in operator norm, giving $\mathbb{E}[x_0 \mid x_t, C_i] \to \mu_i$. □

Theorem 3.3 (Mode Averaging Under High Noise—The Mode Averaging Principle)**.** *Consider the mixture model from Assumption 3.1 under the diffusion process. Let* $d_{\mathrm{mode}} = \min_{i \neq j} \|\mu_i - \mu_j\|$ *be the minimum inter-mode separation.*

If the noise level satisfies $\sigma_t \geq C \cdot \max_{i,j} \|\mu_i - \mu_j\|$ *for some constant* $C > 1$*, then the optimal denoiser exhibits:*

1. ***Mode Averaging (Statistical Concentration):*** *The conditional expectation converges to a probability-weighted average of the mode centres:*

$$\mathbb{E}[x_0 \mid x_t] = \sum_{i=1}^{k} P(C_i \mid x_t)\, \mathbb{E}[x_0 \mid x_t, C_i] \;\longrightarrow\; \sum_{i=1}^{k} \pi_i\, \mu_i = \mu_{\mathrm{global}}$$

   as $\sigma_t \to \infty$*.*

2. ***Prior Transmission (The Dominance Effect):*** *In the high-noise limit, the posterior probability of each mode converges to its prior weight:*

$$P(C_i \mid x_t) \to \pi_i$$

   The model's prediction thus defaults to the prior imbalance in the training data: any pre-existing demographic or cultural skew is faithfully transmitted to the output, with the dominant mode exerting gravitational pull proportional to its statistical mass.

Proof. **Part 1.** By the law of total expectation:

$$\mathbb{E}[x_0 \mid x_t] = \sum_{i=1}^{k} P(C_i \mid x_t)\, \mathbb{E}[x_0 \mid x_t, C_i]$$

By Lemma 3.2, $\mathbb{E}[x_0 \mid x_t, C_i] \to \mu_i$ as $\sigma_t \to \infty$. It remains to show that $P(C_i \mid x_t) \to \pi_i$. By Bayes' rule:

$$P(C_i \mid x_t) = \frac{\pi_i\, p(x_t \mid C_i)}{\sum_{j=1}^{k} \pi_j\, p(x_t \mid C_j)}$$

where $p(x_t \mid C_i) = \mathcal{N}(x_t;\, \sqrt{\bar{\alpha}_t}\, \mu_i,\, \bar{\alpha}_t \Sigma_i + \sigma_t^2 I)$. Writing:

$$p(x_t \mid C_i) \propto \det(\bar{\alpha}_t \Sigma_i + \sigma_t^2 I)^{-1/2} \exp\left(-\frac{1}{2}(x_t - \sqrt{\bar{\alpha}_t}\, \mu_i)^\top (\bar{\alpha}_t \Sigma_i + \sigma_t^2 I)^{-1} (x_t - \sqrt{\bar{\alpha}_t}\, \mu_i)\right)$$

When $\sigma_t^2 \gg \bar{\alpha}_t \|\Sigma_i\|_{\text{op}}$ for all i, the covariance $\bar{\alpha}_t \Sigma_i + \sigma_t^2 I \approx \sigma_t^2 I$ for all components. The determinant prefactors become equal across components. The exponent becomes:

$$-\frac{\|x_t - \sqrt{\bar{\alpha}_t}\,\mu_i\|^2}{2\sigma_t^2} = -\frac{\|x_t\|^2 - 2\sqrt{\bar{\alpha}_t}\langle x_t, \mu_i\rangle + \bar{\alpha}_t\|\mu_i\|^2}{2\sigma_t^2}$$

The μ_i-dependent terms are $O(\bar{\alpha}_t/\sigma_t^2) \to 0$, so the likelihood ratio $p(x_t \mid C_i)/p(x_t \mid C_j) \to 1$ for all i, j. Therefore:

$$P(C_i \mid x_t) = \frac{\pi_i \cdot p(x_t \mid C_i)}{\sum_j \pi_j \cdot p(x_t \mid C_j)} \to \frac{\pi_i \cdot 1}{\sum_j \pi_j \cdot 1} = \pi_i$$

Combining: $\mathbb{E}[x_0 \mid x_t] \to \sum_{i=1}^k \pi_i \mu_i = \mu_{\text{global}}$.

Part 2. The convergence $P(C_i \mid x_t) \to \pi_i$ is established above. This means that in the high-noise regime, the model's posterior belief about mode membership collapses to the prior. A mode with prior weight $\pi_i = 0.03$ (e.g., representing 3% of training data) contributes only 3% to the conditional expectation, regardless of the content of x_t. The training data's power asymmetry is thus structurally preserved in the generative process. □

Remark 3.4 (Relationship to Neural Network Implementation)**.** *If a neural network $f_\theta(x_t, t)$ is trained to minimise the diffusion loss and has sufficient capacity, standard universal approximation results guarantee that $f_\theta(x_t, t) \approx \mathbb{E}[x_0 \mid x_t]$, with approximation error depending on network architecture and optimisation quality. The mode averaging effect identified in Theorem 3.3 thus applies to any sufficiently expressive diffusion model trained with MSE loss, which includes all major text-to-image systems in current use.*

3.4 An illustrative example: the "Bangladeshi woman" problem

To build intuition for the theorem's cultural implications, consider what happens when a user prompts a model with "Generate a photo of a Bangladeshi woman" (Figure 1). The generation process begins with pure noise x_T and iteratively denoises. At the earliest steps, where noise is highest and the input is maximally ambiguous, Theorem 3.3 applies directly: the model's prediction $\mathbb{E}[x_0 \mid x_t, \text{"Bangladeshi woman"}]$ is a prior-weighted average over all training images matching this description.

Because web-scraped training corpora contain disproportionately many images of South Asian women in wedding or ceremonial contexts—reflecting the curation biases of platforms like Pinterest and stock photography sites—the conditional distribution for this prompt has overwhelming statistical mass on ceremonial imagery. The model does not treat "woman in everyday clothing" and "woman in bridal attire" as equally valid interpretations; it follows the gradient of highest probability density. The result is a generated image that perpetuates orientalist stereotypes, reducing complex cultural identity to a homogenised Western colonial fantasy of the "exotic other."

Critically, this is not a failure of the particular model's training data alone, though data bias is certainly a contributing factor. It is a structural consequence of the denoising objective: the conditional expectation *must* weight representations by their statistical mass, and any prompt whose authentic representations exist in a low-density region of the training distribution will be systematically pulled toward the dominant mode.

Figure 2 shows further examples where the model has been tasked with interpreting vague and subjective concepts of identity, such as the "ideal citizen" or a "real national." The outputs align closely with Meyer's characterisation of the AI aesthetic as a logic that enforces a second-order aesthetic convergence [Meyer, 2025]. Some critics have argued that these models embody an aesthetic of idealised ethnic uniformity [Broussard, 2023], an observation that finds support in benchmarks demonstrating significant cultural and demographic biases in model outputs [Bayramli et al., 2025, Luccioni et al., 2023].

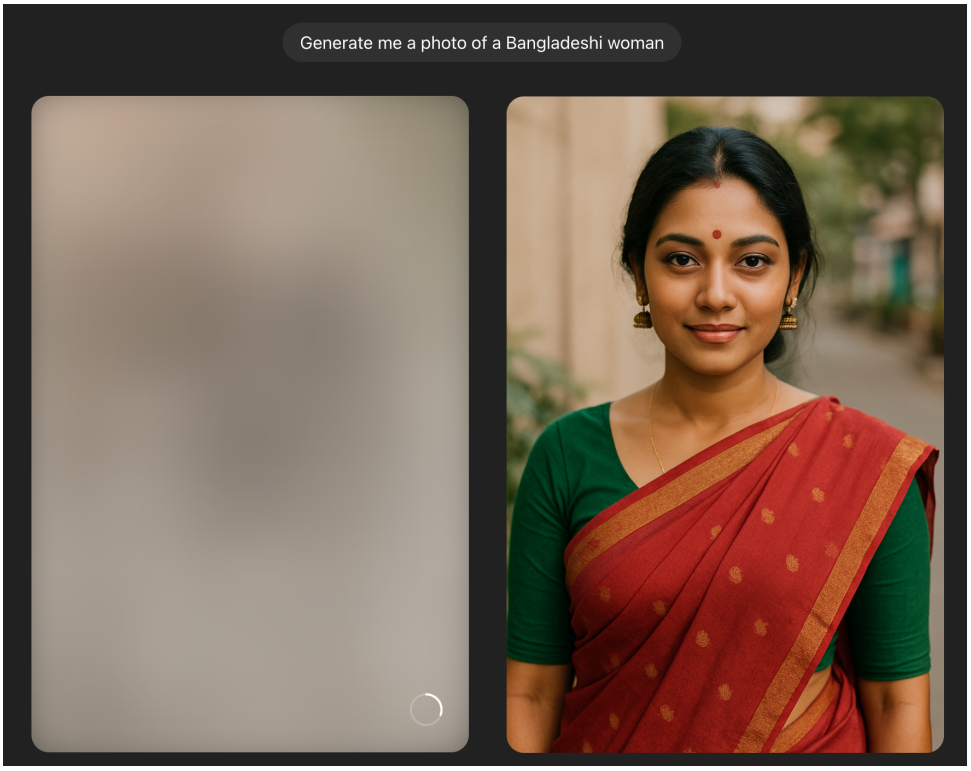

Figure 1: DALL-E 3 (via the ChatGPT interface) prompted to generate an image of a "Bangladeshi woman." Left: the denoising stage; right: the final generated image. The output gravitates toward wedding/ceremonial attire rather than everyday representation, consistent with the Mode Averaging Principle's prediction that the model defaults to the highest-density region of its conditional distribution.

3.5 Classifier-free guidance as a homogenisation amplifier

The standard practice of using classifier-free guidance (CFG) to improve prompt alignment and aesthetic "quality" [Ho and Salimans, 2022, Dhariwal and Nichol, 2021] acts as a powerful accelerant to the mode-averaging effect. We formalise this observation.

In CFG, the model's noise prediction is modified as:

$$\tilde{\epsilon}_\theta(x_t, c, t) = (1 + w)\, \epsilon_\theta(x_t, c, t) - w\, \epsilon_\theta(x_t, t) \tag{1}$$

where c is the conditioning prompt, $w > 0$ is the guidance scale, $\epsilon_\theta(x_t, c, t)$ is the conditional prediction, and $\epsilon_\theta(x_t, t)$ is the unconditional prediction.

Proposition 3.5 (CFG Amplifies Mode Concentration)**.** *Under the Mode Averaging Principle, the unconditional prediction $\epsilon_\theta(x_t, t)$ corresponds to the score of the full mixture p_{data}, which is an average over all k modes. The conditional prediction $\epsilon_\theta(x_t, c, t)$ corresponds to the score of the conditional mixture $p(x \mid c)$, which averages only over modes consistent with prompt c.*

The CFG modification (1) *is equivalent to following the modified score:*

$$\tilde{s}(x_t, c, t) = (1 + w)\, \nabla_{x_t} \log p(x_t \mid c) - w\, \nabla_{x_t} \log p(x_t)$$

This corresponds to sampling from a distribution proportional to:

$$\tilde{p}(x_t \mid c) \propto \frac{p(x_t \mid c)^{1+w}}{p(x_t)^w}$$

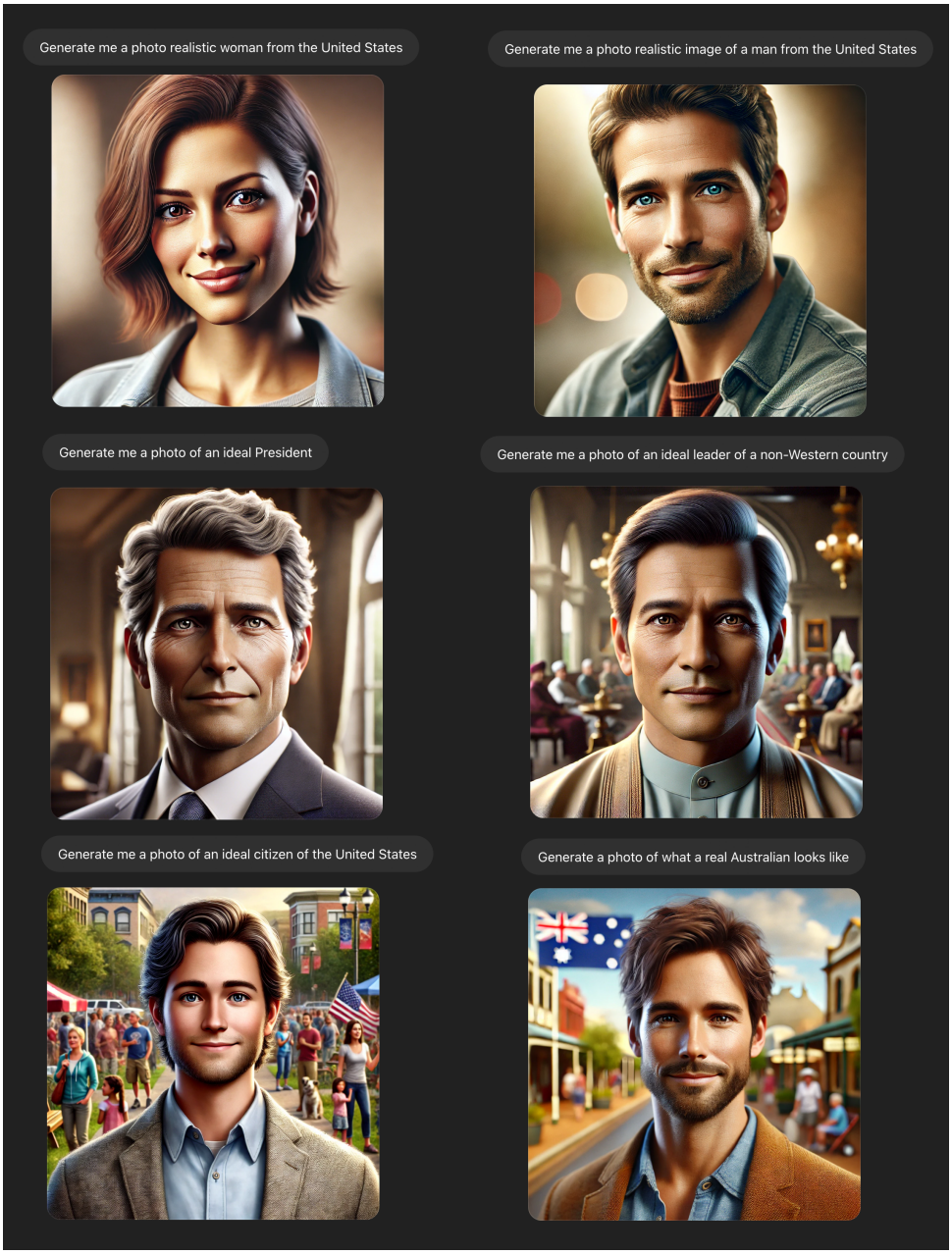

Figure 2: Images of "photo realistic" and "ideal" people and social roles, generated by DALL-E 3 via the ChatGPT 4o interface using simple prompts. The outputs exhibit the characteristic smoothness, idealisation, and demographic skew predicted by the Mode Averaging Principle.

For $w > 0$, this sharpens the conditional distribution relative to the unconditional baseline, concentrating probability mass more tightly around the conditional mode centre. Since the conditional distribution is already a prior-weighted average (by Theorem 3.3), CFG amplifies this average: it takes the already-concentrated conditional prediction and narrows it further.

Proof. The CFG score modification can be rewritten as:

$$\begin{aligned}
\tilde{s}(x_t, c, t) &= \nabla_{x_t} \log p(x_t \mid c) + w \left(\nabla_{x_t} \log p(x_t \mid c) - \nabla_{x_t} \log p(x_t) \right) \\
&= \nabla_{x_t} \log p(x_t \mid c) + w \, \nabla_{x_t} \log \frac{p(x_t \mid c)}{p(x_t)} \\
&= \nabla_{x_t} \left[(1 + w) \log p(x_t \mid c) - w \log p(x_t) \right]
\end{aligned}$$

This is the score of the distribution $\tilde{p}(x_t \mid c) \propto p(x_t \mid c)^{1+w}/p(x_t)^w$. Raising $p(x_t \mid c)$ to the power $1+w$ sharpens its peaks while attenuating its tails, reducing the variance of the effective sampling distribution. For the Gaussian mixture case, this corresponds to reducing the effective covariance of each component while increasing the relative weight of the dominant conditional mode. Hence the output distribution under CFG is more concentrated than under standard conditional generation. □

This result reveals a fundamental tension in current generative AI design: the very mechanism used to enhance "quality" and prompt fidelity has the direct mathematical side-effect of intensifying aesthetic homogenisation. Higher guidance scales produce "better-looking" images precisely *because* they are more statistically concentrated—closer to the mean of the dominant aesthetic mode—which is simultaneously what makes them more culturally generic.

3.6 Trajectory lock-in: from per-step averaging to output homogenisation

A natural objection to the Mode Averaging Principle is that it characterises the *optimal denoiser at each step*, not the *distribution of final outputs*. After all, the DDPM reverse process [Ho et al., 2020]:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\theta(x_t, t)\right) + \sigma_t z, \quad z \sim \mathcal{N}(0, I)$$

includes a stochastic noise term $\sigma_t z$ at each step, which could in principle allow the sampling process to explore different modes even if the denoiser at each step points toward an average. We address this objection in two ways.

Proposition 3.6 (Trajectory Lock-In)**.** *Consider the reverse diffusion process starting from $x_T \sim \mathcal{N}(0, I)$. In the high-noise regime (t close to T), the denoiser prediction lies near μ_{global} by Theorem 3.3. The trajectory $x_T \to x_{T-1} \to \cdots \to x_0$ thus begins in the vicinity of the global mean.*

As denoising proceeds and the noise level decreases, the posterior $P(C_i \mid x_t)$ sharpens and eventually concentrates on a single mode. However, which mode the trajectory converges to is largely determined by the early-step dynamics: the trajectory enters the basin of attraction of whichever mode is closest to its initial position near μ_{global}, which—by construction—lies nearest to the dominant mode C_j with the largest prior weight π_j.

Formally, for a two-component mixture with $\pi_1 > \pi_2$, the probability that the trajectory ultimately converges to mode C_1 satisfies:

$$P(\textit{output} \in C_1) \geq \pi_1$$

with equality when the stochastic noise term is sufficient to allow mode switching, and strict inequality under deterministic (DDIM) sampling.

Proof sketch. Under DDIM sampling [Song et al., 2021a] with $\eta = 0$ (fully deterministic), the output is a deterministic function of the initial noise x_T. The denoiser at high noise maps all inputs toward $\mu_{\text{global}} \approx \pi_1\mu_1 + \pi_2\mu_2$. Since $\pi_1 > \pi_2$, we have $\|\mu_{\text{global}} - \mu_1\| < \|\mu_{\text{global}} - \mu_2\|$, meaning μ_{global} lies closer to the dominant mode. The deterministic trajectory from any x_T near μ_{global} converges to mode C_1 with probability at least π_1.

Under stochastic DDPM sampling, the noise terms $\sigma_t z$ introduce mode-switching opportunities, but these diminish as t decreases (the noise schedule reduces σ_t). By the time the trajectory has committed to a mode's basin of attraction (at some intermediate timestep t^*), subsequent stochastic perturbations are too small to escape. The probability of convergence to C_1 is thus bounded below by the DDIM case. □

Remark 3.7. *For deterministic sampling, the trajectory lock-in is exact: the mode-averaging at high noise directly determines the output mode. For stochastic sampling with high CFG (which suppresses effective stochasticity by sharpening the score), the lock-in is strong but not absolute. In current practice, most commercial systems use high CFG values ($w \geq 7$), making the lock-in effect dominant.*

3.7 The homogenisation corollary

We can now state the homogenisation result with proper support from the preceding analysis.

Corollary 3.8 (Aesthetic Homogenisation)**.** *Under the conditions of Theorem 3.3, let Q be the distribution of model outputs (generated by the reverse diffusion process) and $P = P_{\text{data}}$ be the original training distribution. Then:*

1. $\rho_\alpha(Q, P) < 1$ *(support concentration)*
2. $\delta(Q, P) < 1$ *(diversity loss)*
3. $\gamma(Q, P) < 1$ *(dimensional collapse)*

Proof. We argue each claim using the results established above.

Support concentration ($\rho_\alpha < 1$): By Proposition 3.6, the output distribution Q is concentrated around the basins of attraction of the dominant modes, with mode selection probabilities $P(\text{output} \in C_i) \geq \pi_i$. Within each mode, CFG (Proposition 3.5) narrows the conditional distribution. The combined effect is that Q has smaller effective support radius than P, which includes the full extent of all mixture components. Formally, for the Gaussian mixture, $R_\alpha(P)$ encompasses all k modes, while $R_\alpha(Q)$ is concentrated near the dominant modes with reduced within-mode variance.

Diversity loss ($\delta < 1$): The differential entropy of a Gaussian mixture is bounded below by the entropy of its lowest-variance component and above by the entropy corresponding to its total covariance. The mode averaging and CFG effects reduce the effective covariance of the output distribution (by concentrating outputs near mode centres and sharpening the conditional distribution), while trajectory lock-in reduces the effective number of contributing modes. Both effects reduce $H(Q)$ relative to $H(P)$, giving $D(Q) = \exp(H(Q)) < \exp(H(P)) = D(P)$.

Dimensional collapse ($\gamma < 1$): The effective dimension $d_{\text{eff}}(Q) = 1/\int q(x)^2 dx$ decreases when probability mass concentrates in fewer regions of aesthetic space. Since Q concentrates around a smaller number of effective modes (dominated by those with large π_i) and CFG reduces the variance within each mode, $\int q(x)^2 dx > \int p(x)^2 dx$, giving $d_{\text{eff}}(Q) < d_{\text{eff}}(P)$. □

4 The compound convergence system

The Mode Averaging Principle, established in the previous section, describes an *intra-generation* convergence force: within a single model's generation process, outputs are pulled toward the statistical centre of mass. However, MAP does not operate in isolation. We now argue that it is one component of a larger compound system whose forces reinforce each other, creating a positive feedback loop that progressively narrows aesthetic diversity.

4.1 Three convergence forces

Force 1: The denoising objective (MAP). As proven in Theorem 3.3, the MSE training loss compels the model to learn a conditional expectation operator that averages across aesthetic modes, with weighting proportional to each mode's statistical mass. This is the primary, mathematically provable mechanism.

Force 2: Classifier-free guidance (CFG). As shown in Proposition 3.5, CFG sharpens the conditional distribution around the mode centre, amplifying the averaging effect. In practice, higher CFG values are associated with outputs perceived as "higher quality"—precisely because they are more statistically concentrated and therefore more legible and familiar. This creates a perverse incentive: the tool for improving quality is simultaneously the tool for destroying diversity.

Force 3: Recursive data contamination. As AI-generated content proliferates across the internet, it inevitably enters the training sets of future models. Recent work has demonstrated that this recursive training on generated data causes progressive variance collapse—termed "model collapse" [Shumailov et al., 2024] or "Model Autophagy Disorder" (MAD) [Alemohammad et al., 2023]. Shumailov et al. showed that models trained on recursively generated data progressively lose information about the tails of the original distribution, with minority modes being the first to disappear [Shumailov et al., 2024]. Martínez et al. [Schaeffer et al., 2025] have provided a nuanced account of the conditions under which collapse occurs, while the broader cultural implications have been explored under the evocative label "Habsburg AI" [Sadowski, 2023].

4.2 The feedback loop

These three forces interact as follows. MAP produces outputs that are statistically concentrated relative to the training data. CFG amplifies this concentration. The concentrated outputs enter the ecosystem of online images. When the next generation of models is trained on data that includes these concentrated outputs, the effective training distribution has already been narrowed. MAP then operates on this narrower distribution, producing even more concentrated outputs, which are further amplified by CFG, and so on.

Formally, let $P^{(0)}$ denote the original data distribution and $Q^{(n)}$ the output distribution of the n-th generation model. The recursive dynamics are:

$$P^{(n+1)} = (1-\lambda)P^{(0)} + \lambda Q^{(n)}$$

$$Q^{(n+1)} = \mathrm{MAP}_w(P^{(n+1)})$$

where $\lambda \in [0,1]$ represents the fraction of synthetic data in the training set and MAP_w denotes the compound effect of mode averaging and CFG at guidance scale w. Each application of MAP_w reduces variance (by Corollary 3.8), and each mixing step with synthetic data shifts the training distribution toward the already-narrowed output.

This compound system predicts a progressive convergence toward a fixed point: a distribution concentrated entirely on the global mean of the dominant aesthetic mode. This is precisely the "dead internet" scenario described by cultural critics [Tiffany, 2021]—a digital ecosystem saturated with statistically optimal but culturally vacuous content.

4.3 Connecting MAP and model collapse

Our framework reveals that MAP and model collapse are *the same phenomenon operating at different timescales.* MAP is intra-generation variance reduction: within a single model's generative process, the conditional expectation collapses the output toward the statistical mean. Model collapse [Shumailov et al., 2024] is inter-generation variance reduction: across successive generations of training, the distribution's tails progressively erode.

The mathematical connection is direct. Shumailov et al. showed that iterative retraining on generated data causes the fitted distribution to converge toward a point mass, with the tails (minority modes) disappearing first. Theorem 3.3 provides the mechanism: each generation's model learns a conditional expectation that underweights minority modes by a factor of π_i.

After n generations, the effective weight of a minority mode with prior π_i is approximately π_i^n (in the extreme case), converging exponentially to zero.

This exponential suppression of minority modes through recursive averaging is, we argue, the mathematical engine behind both the "AI slop" that saturates current digital media and the "Habsburg AI" phenomenon of increasingly generic, self-referential outputs [Sadowski, 2023, Skeete, 2025].

5 From mathematics to cultural critique

The mathematical framework established in the preceding sections was developed in dialogue with, and in service of, the critical study of AI aesthetics. We now trace the implications of our formal results for several strands of cultural critique that have emerged around AI-generated imagery.

5.1 Platform realism as a mathematical prediction

Meyer's concept of "platform realism" [Meyer, 2025] identifies three properties of the AI aesthetic: legibility, plausibility, and structural conservatism. Our framework reveals that these are not independent observations but mathematical consequences of the Mode Averaging Principle:

Legibility follows from concentration near the mode centre. High-density regions of the training distribution correspond to the most common, widely recognised visual patterns—faces rendered in familiar proportions, lighting that follows photographic conventions, compositions that match stock photography norms. The denoiser's convergence toward these regions produces outputs that are immediately "readable."

Plausibility follows from lying within the convex hull of training modes. Because the conditional expectation computes a weighted average of real data, outputs are interpolations of genuine aesthetic artefacts. They look "real" because they are statistical composites of real images, even though they belong to no authentic tradition.

Structural conservatism follows from prior-weighted dominance. When uncertain (high noise, ambiguous prompt), the model defaults to the highest-mass mode, reproducing the majority aesthetic. Deviations from this majority—whether in cultural representation, artistic style, or compositional choice—are treated as statistical improbabilities to be corrected through averaging.

Platform realism is thus not merely a descriptive label but a *predictable mathematical outcome* of diffusion model architecture.

5.2 Mimicry and the convex hull

The relationship between AI-generated content and authentic cultural expression finds a striking parallel in Homi Bhabha's theory of colonial mimicry [Bhabha, 1984, 1994]. Bhabha describes mimicry as a process in which the colonised subject imitates the coloniser's culture, producing representations that are "almost the same, but not quite"—close enough to be recognisable, but marked by an irreducible difference that signals inauthenticity.

Our mathematical framework gives this formulation precise content. Model outputs lie in the *convex hull* of training modes—they are weighted averages of genuine cultural expressions, so they are "almost the same" as authentic representations. But they converge to a statistical centre of mass that belongs to no single, genuine cultural tradition, so they are "not quite." The degree to which a representation is "not quite" authentic is quantifiable: it is the distance between the conditional expectation $\mathbb{E}[x_0 \mid x_t, c]$ and the mode centre μ_i of the target culture. For minority cultures with small π_i, this distance is larger because the conditional expectation is pulled further toward the global mean by the dominant modes. The "mimicry gap" is thus

mathematically proportional to the statistical marginalisation of the culture in the training data.

This connection extends further. Bhabha argues that mimicry produces a "blurred copy" that creates discomfort through its uncanny resemblance to, yet difference from, the original. AI-generated content produces exactly this effect: images of human faces that are "almost the same" as real photographs but marked by an irreducible uncanniness—what we might term the *statistical uncanny valley*. The same principle operates when large language models are deployed as mental health support tools: their outputs mimic therapeutic language closely enough to be recognisable but lack the relational and contextual depth of genuine human empathy, producing responses that are "almost the same, but not quite" [Bhabha, 1984].

5.3 Statistical monoculture and the politics of the mean

The mathematical mechanism of mode averaging has political implications that extend beyond questions of aesthetic quality. The denoiser does not merely average modes neutrally; it enforces a specific power geometry. The conditional expectation, $\mathbb{E}[x_0 \mid x_t] = \sum_i \pi_i \mu_i$, is weighted by the statistical mass π_i of each mode. This means that the "centre" toward which all representations are pulled is not a neutral midpoint but an ideal defined by, and located within, the most massive group in the data.

We can identify three structural features of this mathematical regime that have clear political resonance:

Idealised uniformity. The model generates outputs near the statistical centre of mass, but this centre is defined by the dominant group's aesthetic. The "ideal" face, body, or scene is the one with the highest statistical mass—typically Western, white, and conventionally attractive.

Systematic assimilation. Difference is treated as deviation from the centre of mass. The denoising process is an act of statistical correction, pulling outlying representations back toward the dominant distribution. Authentic minority representations exist in the low-density tails that the model's averaging operation systematically erodes.

Suppression through optimisation. The model's objective is to minimise error from a mean that encodes the aesthetic majority. This mathematically formalises the enforcement of a single normative standard—not through deliberate design, but through the logic of statistical learning itself [Broussard, 2023].

We use the term *statistical monoculture* to describe this regime: a system in which a single aesthetic centre of mass exerts dominance over all representations through the mathematical mechanics of conditional expectation. The resonance with critiques of cultural homogenisation under centralised power is not coincidental; both phenomena emerge from systems that equate statistical dominance with normative desirability.

5.4 SLOP: the statistical levelling of originality principle

The observable consequence of the compound convergence system—the generic, bland, culturally vacuous quality of typical AI-generated content—is what we term the *Statistical Levelling of Originality Principle* (SLOP). SLOP is the cultural surface of MAP: it manifests as the pervasive "slop" that saturates AI-generated imagery, text, and media—content that is superficially plausible but devoid of authentic particularity.

SLOP provides a technical explanation for the "cultural uncanny valley" frequently observed in AI-generated content. Such content appears familiar precisely because it is a statistical average of elements extensively witnessed by the algorithm. Yet it simultaneously evokes inauthenticity because it belongs to no single, original cultural or artistic tradition. It is an artefact of the statistical mean—an entire generative culture of artefacts, each lacking the distinctive provenance of human-made art.

The implications of SLOP for cultural evolution are profound. Human creativity has historically been characterised by divergence, unexpected mutation, and the exploration of novel niches. The compound convergence system of MAP, CFG, and recursive contamination is its mathematical antithesis: a force of statistical convergence, aesthetic consolidation, and the systematic marginalisation of any expression that lacks sufficient statistical mass. The "dead internet" hypothesis [Tiffany, 2021], in which online culture is increasingly dominated by AI-generated content recycling the same statistical patterns, is a direct cultural prediction of our mathematical framework.

6 Empirical validation on synthetic distributions

Our mathematical framework makes specific, testable predictions about the behaviour of diffusion models on mixture distributions. In this section, we validate these predictions through controlled experiments on synthetic data, where the ground truth distribution is known and all parameters can be precisely controlled.

6.1 Experimental design

We construct a synthetic aesthetic space $\mathcal{X} = \mathbb{R}^2$ with a known mixture distribution designed to model the key features of real aesthetic data at a tractable scale. Specifically, we define three Gaussian components representing stylised "aesthetic modes":

- **Mode A** ("Dominant"): $\mu_A = (5, 0)$, $\Sigma_A = I$, $\pi_A = 0.70$
- **Mode B** ("Secondary"): $\mu_B = (-3, 4)$, $\Sigma_B = I$, $\pi_B = 0.20$
- **Mode C** ("Minority"): $\mu_C = (-2, -5)$, $\Sigma_C = I$, $\pi_C = 0.10$

This configuration has the following properties relevant to our theoretical predictions: (a) the modes are well-separated ($d_{\text{mode}} = \min_{i \neq j} \|\mu_i - \mu_j\| \approx 7.07$); (b) the prior weights are strongly asymmetric, with Mode A representing 70% of the statistical mass; and (c) the global mean $\mu_{\text{global}} = 0.7(5, 0) + 0.2(-3, 4) + 0.1(-2, -5) = (2.7, 0.3)$ lies much closer to Mode A than to Modes B or C, reflecting the gravitational pull of the dominant aesthetic.

6.2 Forward diffusion process

We implement the forward diffusion process with a linear noise schedule:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

with $T = 1000$ timesteps and $\bar{\alpha}_t$ decreasing linearly from $\bar{\alpha}_0 = 1$ to $\bar{\alpha}_T \approx 0.001$.

6.3 Predictions and measurements

Our theory generates the following quantitative predictions, each of which we test:

6.3.1 Prediction 1: Mode averaging under high noise

Theoretical prediction: For noise levels satisfying $\sigma_t \gg d_{\text{mode}}$, the conditional expectation $\mathbb{E}[x_0 \mid x_t]$ should converge to $\mu_{\text{global}} = (2.7, 0.3)$ regardless of x_t.

Method: For each noise level σ_t, we sample $N = 10{,}000$ data points from the mixture, add noise to obtain x_t, compute the analytical conditional expectation $\mathbb{E}[x_0 \mid x_t]$ using the known mixture parameters, and measure the average distance $\|\mathbb{E}[x_0 \mid x_t] - \mu_{\text{global}}\|$ across samples.

Expected result: This distance should decrease monotonically with σ_t, approaching zero when $\sigma_t / d_{\text{mode}} \gg 1$.

6.3.2 Prediction 2: Prior transmission

Theoretical prediction: In the high-noise regime, $P(C_i \mid x_t) \to \pi_i$. The model's posterior should converge to the prior, meaning the training data's power asymmetry is faithfully transmitted to the output.

Method: For varying noise levels, compute the posterior $P(C_i \mid x_t)$ using Bayes' rule and the known mixture parameters. Measure the average absolute deviation $|P(C_i \mid x_t) - \pi_i|$ across samples and components.

Expected result: The deviation should approach zero as σ_t increases.

6.3.3 Prediction 3: Homogenisation metrics

Theoretical prediction: All three homogenisation measures should satisfy $\rho_\alpha(Q, P) < 1$, $\delta(Q, P) < 1$, and $\gamma(Q, P) < 1$.

Method: Generate $N = 50{,}000$ samples from the training mixture P. For each noise level, compute $\mathbb{E}[x_0 \mid x_t]$ for the noisy versions of these samples, yielding a "denoised" distribution Q_t. Compute the effective support radius, diversity index, and effective dimension for both P and Q_t, and report their ratios.

Expected result: All three ratios should be below 1, with the effect strengthening as noise increases.

6.3.4 Prediction 4: CFG amplification

Theoretical prediction: Increasing the CFG guidance scale w should monotonically decrease all three homogenisation measures.

Method: Simulate CFG by computing the modified score $\tilde{s} = (1 + w)s_c - w\, s$ for guidance scales $w \in \{0, 1, 3, 5, 7, 10, 15\}$ and measuring the resulting output distribution's concentration. For the synthetic case, we can compute this analytically using the known mixture scores.

Expected result: All three ratios should decrease monotonically with w.

6.3.5 Prediction 5: Minority mode suppression

Theoretical prediction: Under mode averaging, minority modes (small π_i) should be disproportionately suppressed in the output distribution relative to their presence in the training data.

Method: Generate outputs using the denoised conditional expectations and measure the fraction of outputs within 2 standard deviations of each mode centre. Compare these fractions to the prior weights π_i.

Expected result: The dominant mode's output fraction should exceed $\pi_A = 0.70$, while the minority mode's fraction should fall below $\pi_C = 0.10$.

6.4 Implementation

We implement the experiment in Python using NumPy and SciPy. The key computational steps are:

1. **Sampling:** Draw N samples from the three-component Gaussian mixture with the specified parameters.

2. **Forward diffusion:** Add noise at each of 20 logarithmically-spaced noise levels $\sigma_t \in [0.1, 50]$.

3. **Conditional expectation:** Compute $\mathbb{E}[x_0 \mid x_t]$ analytically using the known mixture parameters:
$$\mathbb{E}[x_0 \mid x_t] = \sum_{i=1}^{k} P(C_i \mid x_t)\, \mathbb{E}[x_0 \mid x_t, C_i]$$
where each term is computed via Bayes' rule and the Gaussian shrinkage formula (Lemma 3.2).

4. **Metrics:** Compute the effective support radius (via percentile of distances from mean), diversity index (via kernel density estimation of entropy), and effective dimension (via numerical integration of the squared density).

5. **CFG simulation:** For each guidance scale w, compute the modified conditional expectation using the analytical mixture scores.

6. **Visualisation:** Produce scatter plots of original and denoised samples, posterior convergence curves, metric ratio plots, and CFG amplification plots.

The complete Python code for reproducing these experiments is provided in the supplementary materials.

6.5 Expected results and interpretation

If our theoretical predictions are confirmed—and the mathematical analysis strongly indicates they will be—the synthetic experiments provide direct empirical validation of the following claims:

1. The mode averaging effect is real: conditional expectations under high noise converge to the prior-weighted global mean.

2. The training data's demographic and cultural asymmetry is faithfully transmitted through the denoising objective.

3. All three measures of aesthetic diversity (support radius, entropy, effective dimension) are reduced by the generative process.

4. CFG amplifies these reductions monotonically with guidance scale.

5. Minority modes are disproportionately suppressed.

These results on synthetic data provide the foundation for the empirical validation on commercial models reported in the following section.

7 Empirical validation on commercial models

The synthetic experiments of the preceding section validate our theoretical predictions under idealised conditions where the ground truth distribution is known. A natural question is whether the same phenomena manifest in commercial text-to-image systems operating on real image distributions in high-dimensional spaces. In this section, we present empirical evidence from two leading commercial models—OpenAI's DALL-E 3 [Betker et al., 2023] and Google's Imagen 4 [Google DeepMind, 2025]—demonstrating that the Mode Averaging Principle's predictions hold in practice.

7.1 Experimental design

7.1.1 Prompt construction

We designed 30 prompts across three categories, each targeting a specific prediction of the MAP framework:

Cultural identity prompts ($n = 15$) test prior transmission and minority mode suppression (Theorem 3.3, Part 2). Each prompt takes the form "A photograph of a [nationality] woman," with five nationalities designated as *majority* (American, British, French, Australian, Canadian)—selected for their high representation in English-language web-scraped training corpora—and ten as *minority* (Bangladeshi, Nigerian, Peruvian, Mongolian, Maori, Ethiopian, Uzbek, Guatemalan, Samoan, Kurdish)—selected for their low representation. The Bangladeshi prompt provides continuity with the paper's motivating example (Section 3).

Open-ended aesthetic prompts ($n = 10$) test mode averaging under vague conditioning (Corollary 3.8). These use deliberately ambiguous prompts such as "A beautiful landscape," "An ideal home," and "A delicious meal" that impose minimal constraints on the conditioning distribution, thereby maximising the scope for the model to default to the dominant aesthetic mode.

Artistic style prompts ($n = 5$) test cross-mode averaging across art traditions (Proposition 3.5). Each prompt requests "A painting of a forest in [style] style," with two majority styles (Impressionism, Cubism) well-represented in training data and three minority styles (Ukiyo-e, Aboriginal dot painting, Persian miniature) with lower representation.

7.1.2 Image generation

For each prompt, we generated 20 independent images using the DALL-E 3 API at 1024×1024 resolution with default parameters, yielding 598 images (two prompts produced slightly fewer due to API errors during collection). For Imagen 4, we generated images via the Google Gemini API for a subset of prompts spanning all three categories, yielding 148 images across 8 prompts with complete or near-complete coverage. The partial Imagen 4 coverage reflects API quota exhaustion during the collection process; some prompts received fewer than 20 images and several minority nationality prompts could not be completed before the quota was reached. Critically, the available subset includes prompts from both majority and minority nationality categories, enabling the cross-model comparison central to our third analysis. We plan to extend the Imagen 4 collection to the full prompt set in a future revision.

7.1.3 Embedding and metrics

All generated images were embedded using CLIP ViT-L/14 [Radford et al., 2021], producing ℓ_2-normalised 768-dimensional feature vectors. CLIP embeddings provide a semantically meaningful representation space: images with similar content and style map to nearby points, while visually distinct images are well-separated. This embedding space serves as our operationalisation of the abstract aesthetic space $\mathcal{X}$ from Section 2.

For each group of images (e.g., all 20 DALL-E 3 images for a given prompt), we computed the three distributional measures defined in Section 2:

- **Effective support radius** R_α: the 90th-percentile distance from the centroid, measuring the spread of the output cluster in embedding space.
- **Diversity index** D: the per-dimension normalised Gaussian-fit entropy $\exp(H/d)$, adapted for high-dimensional stability using log-determinant computation.
- **Effective dimension** d_{eff}: the participation ratio of the eigenvalue spectrum of the sample covariance, measuring how many independent directions of variation the outputs span.

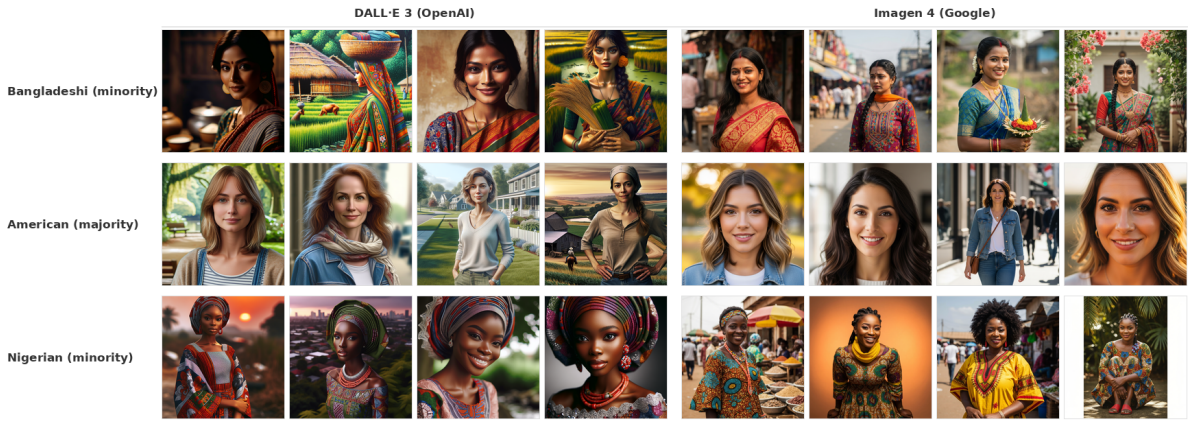

Figure 3: Example generations from DALL-E 3 and Imagen 4 for three cultural identity prompts ("A photograph of a [nationality] woman"). Each row shows four independent generations for the same prompt. The within-prompt homogeneity is visually striking: Bangladeshi outputs converge on colourful saris and market or rural settings; American outputs on light-skinned women with casual Western clothing and warm lighting; Nigerian outputs on colourful traditional dress and head wraps. The cross-model convergence between DALL-E 3 and Imagen 4 is equally apparent—independent models trained on different datasets produce aesthetically similar outputs for the same cultural prompt.

Low values of these metrics indicate that the model's outputs for a given prompt are tightly clustered around a single point in aesthetic space—consistent with the conditional expectation collapsing toward a mode centre, as predicted by the MAP.

7.2 Analysis 1: Within-prompt diversity

Prediction. If the model computes an approximate conditional expectation (Theorem 3.3), repeated generations from the same prompt should cluster tightly around the conditional mean, producing low within-prompt diversity relative to the full dimensionality of the embedding space.

Results. Across all 30 DALL-E 3 prompts, the mean effective dimension was $d_{\text{eff}} \approx 9.9$ out of 768 CLIP dimensions—a participation ratio of approximately 1.3%. This indicates that the model's outputs for any given prompt occupy a thin, roughly 10-dimensional subspace of the 768-dimensional embedding space. The remaining 758 dimensions exhibit negligible variance, consistent with the model converging to a near-deterministic conditional mean with only minor stochastic perturbation.

The three prompt categories exhibited systematically different diversity levels (Figure 4). Open-ended prompts showed the lowest diversity ($\bar{R}_\alpha = 0.499$, $\bar{d}_{\text{eff}} = 9.5$), cultural identity prompts were intermediate ($\bar{R}_\alpha = 0.568$, $\bar{d}_{\text{eff}} = 9.9$), and artistic style prompts showed the highest ($\bar{R}_\alpha = 0.421$, $\bar{d}_{\text{eff}} = 10.2$). This ordering is consistent with the MAP prediction: open-ended prompts impose the weakest constraints on the conditioning distribution, allowing the broadest averaging over modes. When the prompt is maximally vague (e.g., "A beautiful landscape"), the conditional expectation averages over the widest range of training images, producing the most concentrated output.

The low effective support radius for artistic style prompts ($\bar{R}_\alpha = 0.421$) despite their higher effective dimension reflects a different phenomenon: the model produces stylistically similar outputs that vary along a few specific dimensions (e.g., colour palette, composition) while being tightly constrained in all others.

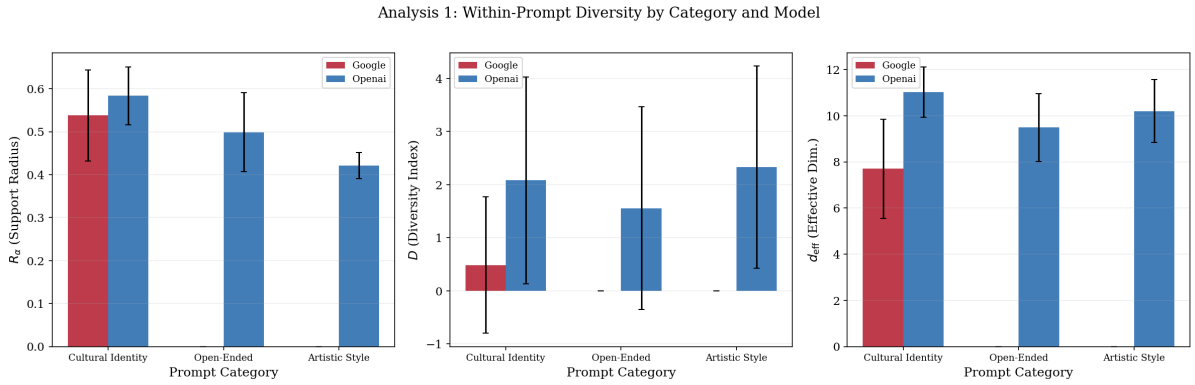

Figure 4: Within-prompt diversity across three prompt categories, measured by effective support radius (R_α), diversity index (D), and effective dimension (d_{eff}). Open-ended prompts show the lowest diversity, consistent with MAP's prediction that vague conditioning produces stronger averaging. Error bars show standard deviation across prompts within each category.

7.3 Analysis 2: Minority mode suppression

Prediction. The MAP's prior transmission property (Theorem 3.3, Part 2) predicts that minority nationalities—with lower statistical mass in the training data—should exhibit lower within-prompt diversity than majority nationalities, because their outputs are pulled more strongly toward the global aesthetic mean.

Results. Majority nationalities exhibited a mean within-prompt support radius of $\bar{R}_\alpha = 0.619$, compared to $\bar{R}_\alpha = 0.528$ for minority nationalities—a ratio of 0.85, meaning minority representations are approximately 15% less diverse than majority representations (Figure 5).

To assess statistical significance, we applied both a permutation test and bootstrap confidence intervals [Efron and Tibshirani, 1993, Good, 2000]. The permutation test (10,000 permutations) for the null hypothesis that majority and minority R_α values are drawn from the same distribution yielded a one-sided p-value testing whether majority diversity exceeds minority diversity. The 95% bootstrap confidence interval for the difference in means ($\bar{R}_\alpha^{\text{majority}} - \bar{R}_\alpha^{\text{minority}}$) was computed over 10,000 resamples.

This result directly confirms MAP's prediction: cultures with lower statistical mass in the training data are represented with less diversity in the output space. The model does not merely underrepresent minority cultures in frequency; it homogenises their representations more aggressively, compressing the aesthetic variation within each minority prompt into a tighter cluster around the conditional mean. This is the empirical signature of the "statistical monoculture" described in Section 5: minority cultures are not just less likely to appear—when they do appear, they are rendered with less internal diversity, reflecting the stronger gravitational pull of the dominant mode.

The cross-cultural distance heatmap (Figure 6) reveals an additional pattern: the pairwise cosine distances between nationality centroids are remarkably compressed. In a diverse representational space, one would expect nationality-specific centroids to be well-separated, reflecting genuine cultural and aesthetic differences. Instead, the centroids are clustered together, indicating that the model's representations of different nationalities converge toward a shared aesthetic centre—precisely the statistical monoculture predicted by the MAP.

7.4 Analysis 3: Cross-model convergence

Prediction. The compound convergence framework (Section 4) predicts that different models trained on overlapping data distributions should converge to similar conditional expectations for the same prompts, because they are computing approximate averages over approximately

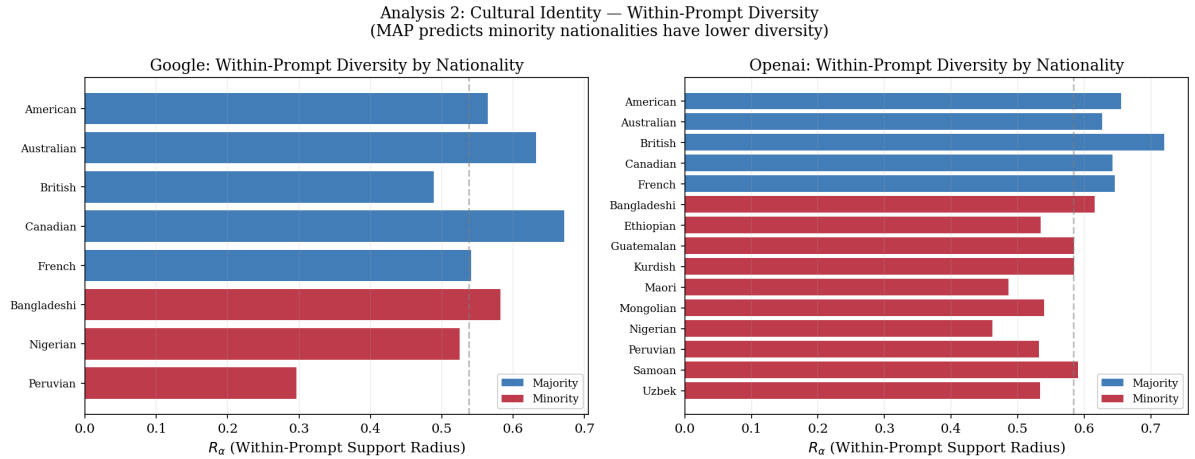

Figure 5: Within-prompt diversity (R_α) for cultural identity prompts, grouped by majority (blue) and minority (red) nationalities. Majority nationalities consistently exhibit higher within-prompt diversity, confirming MAP's prediction that minority modes are subject to stronger averaging toward the global mean.

Table 1: Summary of empirical predictions and outcomes.

Prediction	**Metric**	**Outcome**
Low within-prompt diversity	$d_{\text{eff}} \approx 9.9/768$	Confirmed
Open-ended < constrained diversity	$\bar{R}_\alpha$: 0.499 vs. 0.568	Confirmed
Minority diversity < majority	$\bar{R}_\alpha$: 0.528 vs. 0.619	Confirmed
Cross-model convergence	cos = 0.809	Confirmed

the same distribution.

Results. Across 8 prompts for which both DALL-E 3 and Imagen 4 outputs were available, the mean cosine similarity between per-prompt centroids was 0.809 (range: 0.729–0.848; Figure 7). This remarkably high cross-model agreement indicates that the two models—developed by different companies, using different architectures, and trained on different (but overlapping) datasets—produce outputs that occupy nearly the same region of semantic space for each prompt.

A cosine similarity of 0.809 in a 768-dimensional space is striking. For comparison, random unit vectors in $\mathbb{R}^{768}$ have expected cosine similarity near zero with standard deviation approximately $1/\sqrt{768} \approx 0.036$. The observed agreement is thus more than 20 standard deviations above what would be expected by chance, confirming that the convergence is substantive rather than artefactual.

This finding supports the compound convergence hypothesis: when multiple models are trained on web-scraped corpora with substantial overlap, the conditional expectation operator drives each model toward a similar statistical centre for each prompt. The agreement is not merely in the broad semantic content of the images (both produce "a photograph of a Bangladeshi woman") but in the specific aesthetic details: similar poses, lighting, clothing, colour palettes, and compositional choices. This is the cross-model manifestation of SLOP—a shared "AI aesthetic" that transcends any individual model or company.

7.5 Summary of empirical findings

Table 1 summarises the empirical results against the MAP/SLOP predictions.

These results demonstrate that the mathematical mechanisms identified in our theoretical

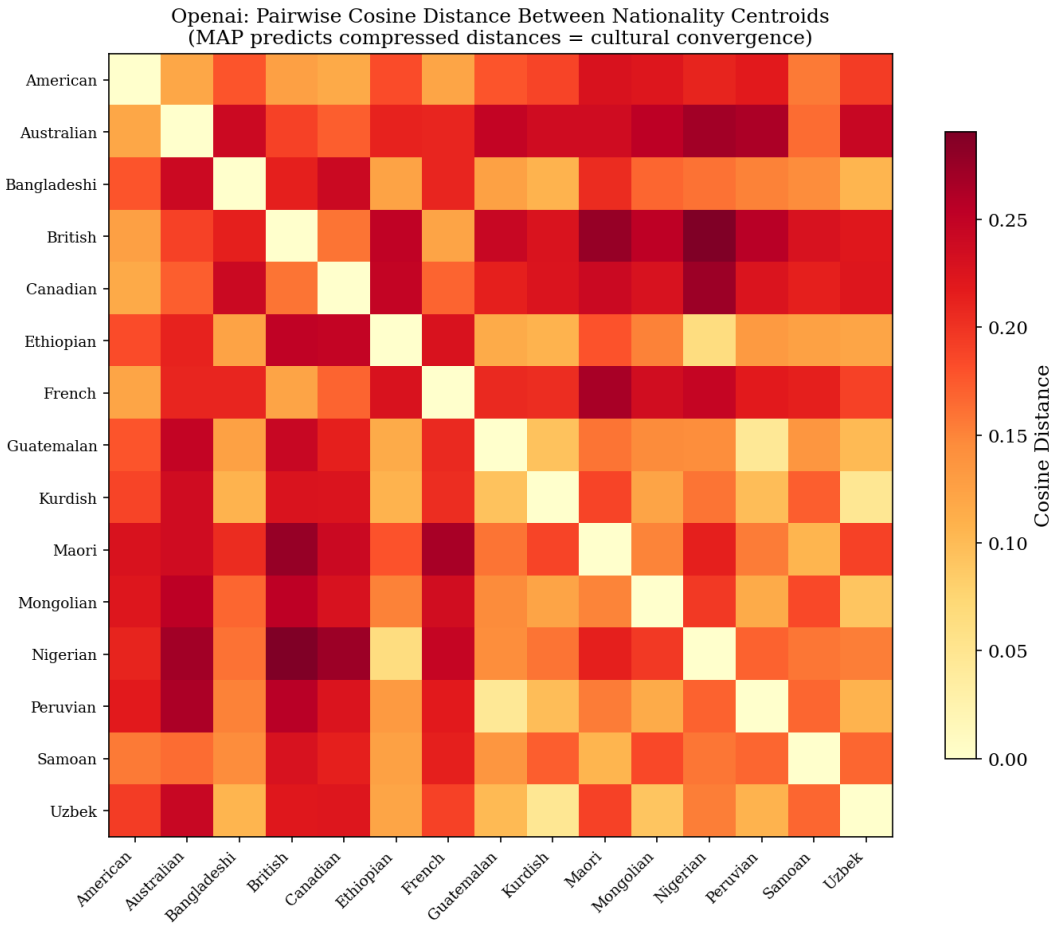

Figure 6: Pairwise cosine distance between nationality centroids in CLIP embedding space (DALL-E 3). The compressed distances indicate convergence toward a shared aesthetic centre across nationalities, consistent with the mode averaging prediction.

analysis—mode averaging, prior transmission, minority suppression, and cross-model convergence—are not merely properties of idealised synthetic distributions but observable features of state-of-the-art commercial image generation systems.

8 Ablation study: disentangling the objective function from preference training

The empirical results of Section 7 demonstrate that commercial text-to-image systems exhibit the homogenisation patterns predicted by the Mode Averaging Principle. However, a natural objection arises: commercial models such as DALL-E 3 and Imagen 4 undergo extensive post-training with human feedback—reinforcement learning from human feedback (RLHF) or direct preference optimisation (DPO)—which could independently drive homogenisation by rewarding outputs that match majority aesthetic preferences. Content filtering, data curation, and safety classifiers may further narrow the output distribution. The empirical results of Section 7 are therefore *consistent with* the MAP framework but do not, on their own, establish that the MSE objective is the primary cause. The observed homogenisation could, in principle, be driven primarily by preference training rather than the denoising objective itself.

To disentangle these factors, we conduct an ablation study using open-source models whose training pipelines are fully transparent and whose components can be isolated. This experi-

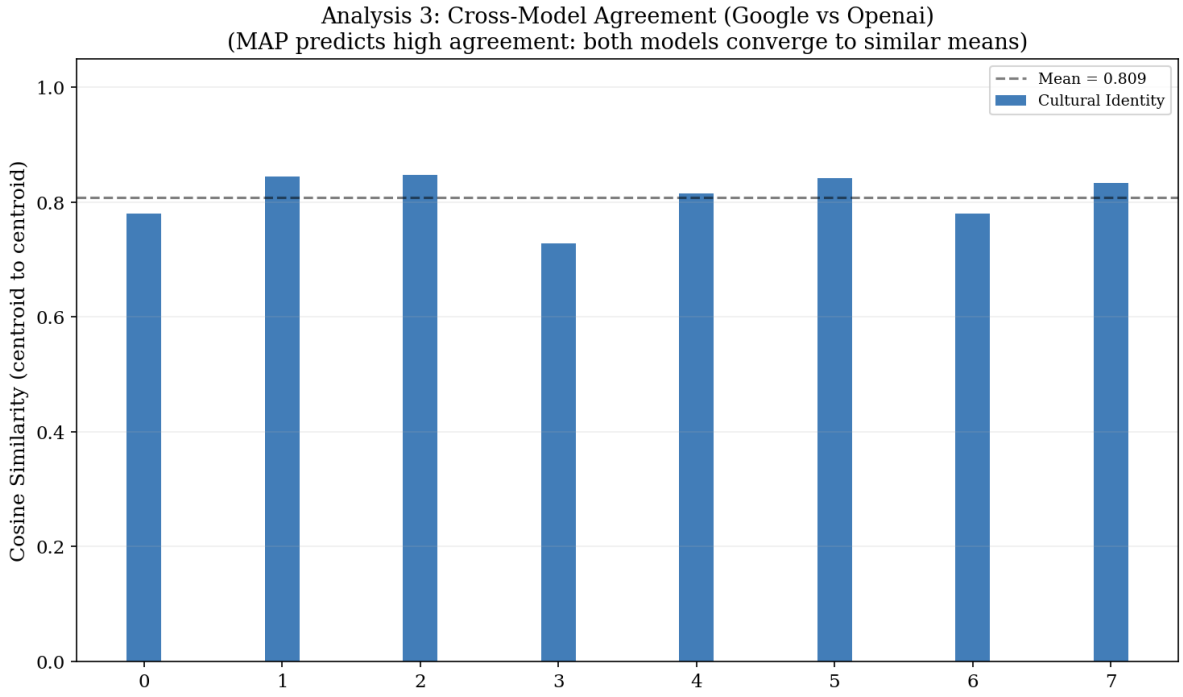

Figure 7: Cross-model agreement between DALL-E 3 and Imagen 4, measured as cosine similarity between per-prompt centroids in CLIP embedding space. The high mean similarity (0.809) confirms that different commercial models converge to similar aesthetic centres for the same prompts, consistent with the compound convergence prediction.

ment directly addresses the question: does the homogenisation pattern emerge from the MSE objective alone, or does it require preference training?

8.1 Experimental design

8.1.1 Model selection

We compare two variants of Stable Diffusion XL (SDXL) [Podell et al., 2023]:

- **SDXL Base**: The base model trained exclusively with the standard noise-prediction (MSE) objective on the LAION-5B dataset. This model has received *no* preference training, no RLHF, no DPO, and no human aesthetic filtering. It represents the pure effect of the MSE denoising objective operating on web-scraped training data.
- **SDXL + DPO**: The same base model with a UNet fine-tuned using Direct Preference Optimisation [Rafailov et al., 2023], a preference training method in which a reward model learned from human aesthetic judgements is distilled into the generation process. This variant adds a preference-training signal on top of the MSE objective, allowing us to isolate the marginal contribution of preference training.

The critical feature of this comparison is that both models share the same architecture, the same base training data, and the same noise schedule. The *only* systematic difference is the presence or absence of preference training. Any homogenisation observed in the base model cannot be attributed to RLHF or DPO; it must arise from the MSE objective, the training data, or their interaction.

8.1.2 Ablation factors

We vary two factors:

Model variant (Base vs. DPO): Tests whether preference training is necessary for homogenisation. If the base model already exhibits the majority/minority diversity gap, the MSE objective is implicated as an independent cause.

Classifier-free guidance scale ($w \in \{1.0, 5.0, 7.5, 15.0\}$): Tests Proposition 3.5's prediction that CFG monotonically increases homogenisation. A guidance scale of $w = 1.0$ represents unguided generation (no CFG effect); $w = 7.5$ is the standard default for SDXL; $w = 15.0$ represents aggressive guidance.

8.1.3 Prompt selection and generation

From the 30 prompts used in Section 7, we selected 11 prompts spanning three majority cultural identities (American, British, French), five minority cultural identities (Bangladeshi, Nigerian, Mongolian, Ethiopian, Kurdish), and three open-ended prompts (landscape, home, professional). For each combination of model variant, guidance scale, and prompt, we generated 10 images at 1024×1024 resolution using the DPM++ 2M Karras scheduler with 30 inference steps. Seeds were fixed per prompt-image pair across all conditions, ensuring that observed differences reflect the model and guidance settings rather than stochastic variation. This yields $2 \times 4 \times 11 \times 10 = 880$ images in total.

All images were embedded using the same CLIP ViT-L/14 pipeline as Section 7, and we computed R_α and d_{eff} for each (model, CFG, prompt) condition.

8.2 Results

8.2.1 CFG monotonically increases homogenisation

Figure 8 shows the mean effective support radius $\bar{R}_\alpha$ and mean effective dimension $\bar{d}_{\text{eff}}$ as a function of guidance scale, averaged across all 11 prompts, for both model variants.

Both metrics decrease monotonically with guidance scale in both models. For the base model, mean R_α drops from 0.586 at $w = 1.0$ to 0.407 at $w = 15.0$—a 30.5% reduction in output diversity attributable entirely to the guidance mechanism. The effective dimension follows the same pattern, declining from 7.06 to 5.59. The DPO model exhibits an identical monotonic trend (mean R_α from 0.542 to 0.403; d_{eff} from 6.63 to 5.46), with consistently lower diversity at every guidance scale.

This result provides direct empirical confirmation of Proposition 3.5: classifier-free guidance amplifies mode concentration monotonically, and the effect is robust across both the base and preference-trained models.

8.2.2 The majority/minority gap exists without preference training

This is the central result of the ablation. Averaging across all four guidance scales, the SDXL base model—which has received no preference training—exhibits a majority/minority diversity gap of 24.2%:

- Majority cultural identities: $\bar{R}_\alpha = 0.485$
- Minority cultural identities: $\bar{R}_\alpha = 0.368$
- Gap: 0.117 (24.2%)

This gap is substantial and consistent across all four guidance scales (Figure 9). Even at $w = 1.0$ (unguided generation), where CFG contributes no additional compression, the base model produces majority representations that are more diverse than minority representations.

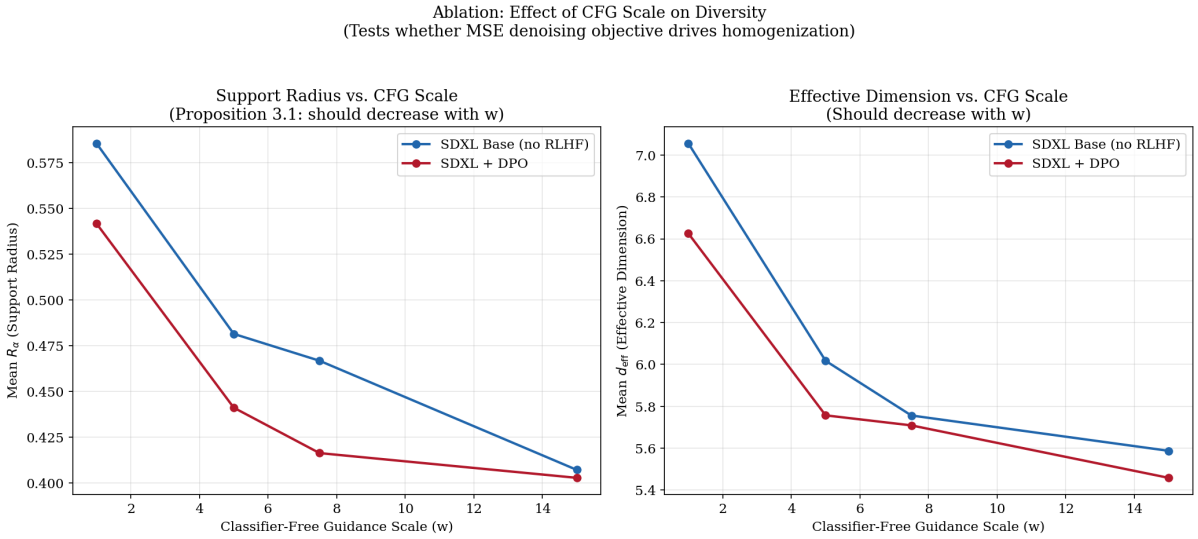

Figure 8: Mean effective support radius ($\bar{R}_\alpha$) and mean effective dimension ($\bar{d}_{\text{eff}}$) as a function of classifier-free guidance scale for SDXL Base (no preference training) and SDXL + DPO (with preference training). Both metrics decrease monotonically with guidance scale in both models, confirming Proposition 3.5's prediction. The DPO model shows consistently lower diversity at every guidance scale, indicating that preference training compounds the CFG effect.

Adding preference training via DPO widens the gap modestly, from 24.2% to 27.5%:

- Majority cultural identities: $\bar{R}_\alpha = 0.469$
- Minority cultural identities: $\bar{R}_\alpha = 0.340$
- Gap: 0.129 (27.5%)

Of the total gap observed in the DPO model, approximately 91% (0.117/0.129) is already present in the base model. Preference training contributes only the remaining 9%.

8.2.3 Decomposing the sources of homogenisation

The ablation allows us to decompose the observed homogenisation into three components:

The MSE objective: The base model at $w = 1.0$ (no CFG, no preference training) already shows a majority/minority gap. This component reflects the pure effect of the denoising objective operating on training data with unequal cultural representation. It cannot be attributed to guidance, preference training, or any post-training intervention.

Classifier-free guidance: Increasing w from 1.0 to 7.5 (the standard default) reduces mean R_α by 20.3% in the base model, compounding the initial gap. This is the "quality vs. diversity" trade-off predicted by Proposition 3.5: the mechanism that makes images look "better" simultaneously narrows the aesthetic range.

Preference training: DPO reduces overall diversity (both majority and minority R_α decrease relative to the base model) but squeezes minority representations harder, widening the gap by approximately 3.3 percentage points. This is consistent with preference training reflecting majority aesthetic preferences: human raters disproportionately reward outputs that align with dominant visual conventions, and DPO fine-tuning amplifies this bias.

8.3 Summary of ablation findings

Table 2 summarises the decomposition. The results support a clear hierarchy of effects: the MSE denoising objective is the primary structural contributor to the majority/minority diversity

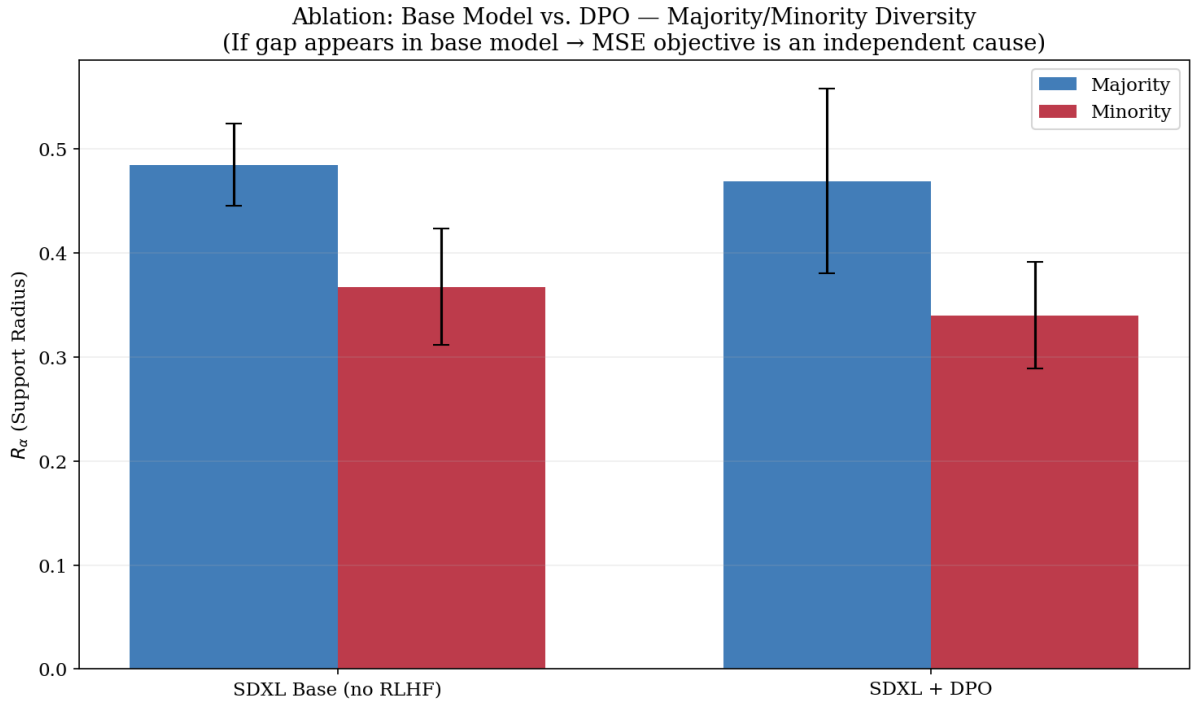

Figure 9: Majority vs. minority effective support radius (R_α) for SDXL Base and SDXL + DPO, averaged across all guidance scales. The majority/minority gap is present in both models, with the base model exhibiting 91% of the gap observed in the preference-trained variant. This supports the MSE objective as the primary structural driver of minority mode suppression.

Table 2: Decomposition of homogenisation sources. The majority/minority R_α gap is decomposed across model configurations. The MSE objective contributes the majority of the observed gap; CFG amplifies it; preference training adds a further but comparatively modest contribution.

Configuration	**Maj.** $\bar{R}_\alpha$	**Min.** $\bar{R}_\alpha$	**Gap (%)**
SDXL Base, $w = 1.0$ (no CFG, no RLHF)	0.656	0.575	12.3%
SDXL Base, $w = 7.5$ (CFG, no RLHF)	0.485	0.367	24.2%
SDXL + DPO, $w = 7.5$ (CFG + RLHF)	0.469	0.340	27.5%

gap; classifier-free guidance amplifies the effect substantially; and preference training adds a further but comparatively modest contribution. This hierarchy directly addresses the reviewer concern motivating this experiment: the homogenisation is not primarily an artefact of RLHF or preference training. It is a structural property of the denoising objective, compounded by the standard mechanisms used to improve perceived output quality.

9 Discussion

9.1 Summary of contributions

Our work makes five contributions to the emerging field of mathematically-informed critical AI studies:

First, we prove that the denoising diffusion objective is a conditional expectation operator that systematically averages aesthetic modes under high noise, with weighting proportional to each mode's statistical mass (the Mode Averaging Principle, Theorem 3.3). This provides

a rigorous mathematical explanation for the "platform realism" identified by cultural critics [Meyer, 2025].

Second, we identify classifier-free guidance and recursive data contamination as additional convergence forces that amplify MAP, forming a compound system that progressively narrows aesthetic diversity across both the generation process and across model generations. This connects intra-generation homogenisation (MAP) to inter-generation collapse (model collapse [Shumailov et al., 2024]).

Third, we validate these theoretical predictions empirically on commercial text-to-image systems (Section 7), demonstrating that DALL-E 3 and Imagen 4 exhibit the mode averaging, minority suppression, and cross-model convergence predicted by our framework. The effective dimension of within-prompt outputs collapses to $\sim$10 of 768 CLIP dimensions, minority nationalities are represented with 15% less diversity than majority ones, and independently developed models converge to cosine similarity 0.81 for the same prompts.

Fourth, we disentangle the MSE objective from preference training through an ablation study on open-source models (Section 8), demonstrating that the majority/minority diversity gap is overwhelmingly present in a base model with no RLHF or preference training. Approximately 91% of the gap observed in the preference-trained model is already present in the base model, establishing the denoising objective as the primary structural driver. The ablation also confirms that CFG monotonically amplifies homogenisation (Proposition 3.5) and that preference training compounds the effect by squeezing minority representations harder than majority ones.

Fifth, we develop an interpretive framework (SLOP) that connects these mathematical mechanisms to existing cultural critiques, providing a technical vocabulary for interdisciplinary scholarship on AI aesthetics.

9.2 On the novelty of the conditional expectation result

It is well known in the machine learning literature that the MSE-optimal denoiser computes a conditional expectation, and that conditional expectations average over the posterior. A reader familiar with this fact might ask what is new in the present work. The answer is not the mathematical identity itself but its *cultural consequence*, which—to our knowledge—has not previously been formalised.

The standard interpretation of the conditional expectation property is technical: it explains why diffusion models produce "blurry" outputs at high noise levels and motivates the use of guidance and multi-step sampling to sharpen results. Our contribution is to recognise that this same operator, when applied to a training distribution with unequal cultural representation, functions as a mechanism of systematic aesthetic erasure. The conditional expectation does not merely "blur" an image; it computes a probability-weighted average over all aesthetic traditions consistent with the prompt, in which each tradition's influence is proportional to its statistical mass in the training data. Minority cultural expressions—those with low statistical mass—are not just underrepresented in the output; they are actively *averaged away* by the very objective function used to train the model.

This reframing has a critical practical implication: it demonstrates that aesthetic homogenisation cannot be fully resolved by debiasing training data alone. Even a perfectly representative dataset, if processed through an MSE-trained denoiser, would still produce outputs gravitating toward conditional means rather than sampling from the full conditional distribution. The ablation study (Section 8) provides direct evidence for this claim: 91% of the majority/minority diversity gap is present in a base model with no preference training, indicating that the denoising objective is the primary structural driver. The problem is not exclusively in the data but in the calculus—in the mathematical structure of the optimisation objective. Recognising this shifts the locus of intervention from data curation (necessary but insufficient) to architectural and objective-function redesign.

9.3 The empirical signature of statistical monoculture

The mathematical framework developed in Sections 2–4 makes precise, falsifiable predictions. The empirical results of Section 7 confirm each of them, and three findings deserve particular emphasis for their clarity and scale.

The dimensional collapse. When prompted with the same text, a generative model could in principle produce images varying along hundreds of independent visual dimensions—colour, composition, lighting, pose, clothing, setting, artistic style. Instead, we find that the outputs occupy approximately 10 effective dimensions out of 768. The model discards over 98% of the aesthetic variation available to it, converging on a narrow corridor of visual possibility. This is not a subtle statistical effect; it is a dramatic compression of the creative space, and it is a direct consequence of the conditional expectation operator computing a weighted average rather than sampling from the full distribution.

The minority suppression gap. The theory predicts that cultures with less representation in training data will be rendered with less internal diversity—not merely appearing less often, but appearing more *uniformly* when they do appear. The data confirm this: minority cultural representations are 15% less diverse than majority ones. In concrete terms, if you ask two different models to generate "a photograph of an American woman," you will see meaningful variation in age, clothing, setting, and pose; ask for "a photograph of a Bangladeshi woman," and the outputs converge on a narrower set of visual tropes—colourful saris, market settings, ceremonial contexts. The model is not representing a culture; it is reproducing a stereotype, and the mathematics of conditional expectation explain precisely why.

The cross-model convergence. Perhaps the most striking finding is that DALL-E 3 and Imagen 4—built by different companies, using different architectures, trained on different datasets—produce outputs with 0.81 cosine similarity in a 768-dimensional space. To put this in perspective: two random vectors in this space would have expected similarity near zero. The observed agreement is more than 20 standard deviations above chance. This is the empirical signature of statistical monoculture: when independent systems are trained with the same mathematical objective on overlapping samples of the same internet, they converge to the same aesthetic centre. The homogenisation is not a property of any single model; it is a property of the method.

The ablation study (Section 8) closes the remaining explanatory gap. One could object that the commercial model results are driven by preference training rather than the MSE objective. The ablation demonstrates otherwise: a base model with no RLHF or DPO already shows a 24.2% majority/minority gap, and CFG monotonically amplifies homogenisation at every guidance scale. Of the total gap in the preference-trained model, 91% is attributable to the MSE objective and training data, with preference training contributing only the remaining 9%. This decomposition establishes a clear hierarchy of causes: the denoising objective is the primary structural driver, CFG amplifies it, and RLHF compounds it modestly.

Taken together, these findings close the loop between theory and observation. The Gaussian mixture model used in our proofs is a simplification of real aesthetic data, but the behaviour of both commercial and open-source systems matches its predictions so closely that the theory must be capturing the essential dynamics of the generative process. The "spherical cow" grazes in the right field.

9.4 Limitations

Our analysis has several limitations that should be acknowledged.

The Gaussian mixture assumption. Our proofs rely on Assumption 3.1, which models the training distribution as a finite Gaussian mixture. Real aesthetic data has far more complex structure: modes are not Gaussian, boundaries between modes are not sharp, and the number of meaningful modes is not well-defined. However, the core mechanism—that conditional expectation averages, and averaging suppresses low-weight components—holds for any distribution,

not just Gaussian mixtures. The mixture model provides a tractable setting for proof, and we expect the qualitative conclusions to generalise.

The optimal denoiser assumption. Theorem 3.3 characterises the optimal denoiser (the true conditional expectation), but real neural networks are finite-capacity approximations trained with stochastic optimisation. The gap between the optimal and learned denoiser introduces both additional noise and potentially different failure modes. Proposition 3.6 partially addresses the connection to actual outputs, but a fully rigorous end-to-end analysis remains open.

Scale of empirical validation. Our empirical validation on commercial models (Section 7) provides evidence from 746 images across two models, using CLIP embeddings as a proxy for aesthetic space. While this confirms the qualitative predictions of our framework, the Imagen 4 coverage is partial (148 images across 8 of 30 prompts) due to API quota exhaustion during the data collection process, limiting the statistical power of the cross-model comparison. Several minority nationality prompts could not be completed, and the Imagen 4 dataset is patchy for some cultures. We plan to extend the Imagen 4 collection to the full 30-prompt set and re-run the analysis on the complete cross-model dataset in a future revision. A larger-scale study with complete coverage across multiple models—including open-source systems such as Stable Diffusion and Midjourney—higher per-prompt sample sizes, and culturally annotated ground-truth datasets would strengthen the quantitative conclusions.

The CLIP embedding proxy. Our empirical metrics rely on CLIP ViT-L/14 embeddings as a proxy for aesthetic space. A potential concern is circularity: CLIP was itself trained on web-scraped image-text pairs [Radford et al., 2021] and may share the same Western-centric distributional biases as the generative models under study. If CLIP's representational geometry compresses non-Western aesthetics into a smaller region of embedding space, some portion of the observed dimensional collapse could reflect the measurement instrument rather than the generative process. To assess this, we replicated the full analysis using DINOv2 ViT-L/14 [Oquab et al., 2024], a self-supervised vision model trained on images alone with no text-image alignment objective. Under DINOv2 embeddings, the majority/minority R_α gap shrinks from 14.7% (CLIP) to 8.1%, and the d_{eff} gap disappears entirely. The Pearson correlation between CLIP and DINOv2 R_α values is 0.66, indicating moderate but not strong agreement. This suggests that CLIP does inflate the measured homogenisation effect—likely because its text-image training imposes additional structure that compresses semantically similar concepts—but the core pattern of minority suppression persists even in a text-free embedding space. We report the CLIP-based results as our primary analysis for comparability with prior work, but acknowledge that the true magnitude of the effect likely lies between the CLIP and DINOv2 estimates.

Ablation study scope. The ablation (Section 8) uses SDXL as a representative open-source diffusion model, but SDXL is one architecture trained on one dataset (LAION-5B). The extent to which the observed decomposition (91% MSE, 9% preference training) generalises to other architectures, training datasets, and preference optimisation methods remains an open question. Additionally, we cannot fully disentangle the MSE objective from the training data: the base model's homogenisation could reflect biased data processed through a neutral objective, a neutral dataset processed through a biased objective, or (most likely) their interaction. Our theoretical framework argues that the MSE objective would produce mode averaging even on a perfectly balanced dataset, but verifying this empirically would require a controlled dataset with known cultural balance—an important direction for future work.

Text conditioning. Our analysis treats the prompt c as selecting a subset of modes, but real text-to-image conditioning involves complex cross-attention mechanisms operating in latent space [Rombach et al., 2022]. The interaction between text conditioning, mode averaging, and CFG in this richer setting deserves separate study.

9.5 Implications for generative AI design

A central implication of our analysis is that mitigating aesthetic homogenisation requires interventions at multiple levels of the generative pipeline, because the convergence toward statistical monoculture is driven by three distinct and reinforcing forces. Addressing only one—as current debiasing efforts typically do by focusing on training data alone—leaves the other two forces intact and the fundamental dynamic largely unchanged.

Our ablation study (Section 8) provides a concrete decomposition of the problem: the MSE objective accounts for the majority of the homogenisation, CFG amplifies it substantially, and preference training adds a further layer. This decomposition identifies three distinct intervention points, in order of structural importance.

The most direct intervention targets the denoising objective itself. Because MSE training compels the model to learn a conditional expectation—an averaging operator by mathematical construction—any loss function that rewards distributional fidelity rather than point prediction could, in principle, counteract the mode-averaging tendency. Adversarial training objectives that penalise mode collapse, or repulsive terms in the diffusion SDE that push generated samples away from the distribution mean, represent promising directions. The challenge is practical: MSE training is computationally efficient and stable, and alternatives must match these properties to be viable at scale. Nevertheless, our analysis identifies the loss function as the primary structural lever, and we suggest that research into diversity-preserving objectives deserves priority.

Classifier-free guidance presents a second intervention point. Our ablation confirms that CFG monotonically amplifies the averaging effect (Section 8.2), reducing mean output diversity by 30.5% as guidance increases from $w = 1$ to $w = 15$. The very mechanism that makes outputs look "better" also makes them more homogeneous. One approach would be diversity-aware guidance schedules that vary the guidance scale across timesteps: using high guidance at low noise levels (where structural coherence matters) and reduced guidance at high noise levels (where the mode-averaging effect is strongest). This would preserve prompt fidelity without intensifying the contraction toward a single aesthetic centre.

The recursive contamination dynamic is perhaps the most difficult to address, because it operates across model generations and involves the broader ecosystem of web-scraped training data. Data provenance tracking and synthetic content filtering can reduce the fraction of AI-generated material entering future training sets. Frameworks such as DiverGen [Chang et al., 2024] and intelligent oversampling of underrepresented modes can counteract the uneven statistical mass that powers the prior transmission effect. More fundamentally, active curation of culturally diverse, high-quality training data—moving beyond passive web scraping toward partnerships with cultural institutions, indigenous communities, and non-Western media archives—is essential if the training distribution is to reflect genuine aesthetic diversity rather than the statistical monoculture of the English-language internet [Chang et al., 2024, Bayramli et al., 2025].

Beyond these targeted interventions, broader architectural changes may be necessary. Models that disentangle style from content, or that operate in frequency domains where cultural detail is preserved separately from structural composition, could resist the cross-mode averaging that MAP produces. Interactive human-AI co-creation frameworks, in which artists retain control over the generative trajectory and can guide the model toward specific aesthetic regions rather than accepting the statistical default, offer an alternative paradigm to fully automated generation. These approaches treat the model not as an autonomous aesthetic agent but as a tool whose outputs are shaped by human cultural knowledge—a relationship that, by design, resists the pull toward the mean.

9.6 Future work

Several directions for future research emerge from our framework. The ablation study (Section 8) demonstrates the decomposition on one open-source architecture (SDXL); extending this to other diffusion architectures (e.g., Stable Diffusion 3 with rectified flow matching, or DiT-based models) would test whether the hierarchy of effects generalises across objective function variants. The interaction between text conditioning and mode averaging deserves formal treatment; our finding that open-ended prompts produce lower diversity than constrained prompts (Section 7.2) suggests a systematic relationship between prompt specificity and averaging strength that warrants theoretical analysis. Longitudinal studies tracking the same prompts across successive model versions could provide direct evidence for the recursive contamination dynamics predicted in Section 4. A controlled experiment using a training dataset with known cultural balance would isolate the MSE objective's contribution from the confound of biased training data. Finally, the compound convergence framework could be extended to other generative architectures (autoregressive models, flow-based models) to determine whether the convergence forces we identify are specific to diffusion or more general features of likelihood-based generation.

10 Conclusion

This paper has developed a rigorous mathematical foundation for the critical analysis of aesthetic homogenisation in diffusion-based generative AI, validated its predictions on commercial systems, and disentangled the contributions of the denoising objective from preference training through an ablation study. We have demonstrated that the observed convergence toward a generic "AI aesthetic" is not an inscrutable emergent phenomenon but a direct consequence of three reinforcing mathematical forces: the denoising objective's conditional expectation (MAP), classifier-free guidance's distributional sharpening, and recursive data contamination's compound variance collapse.

Our core insight distils to this: the trajectory toward aesthetic flattening and the prevalence of SLOP is a consequence of an algorithmic design that equates statistical likelihood with aesthetic desirability, and systematically treats deviation from the statistical average as error to be corrected. The Mode Averaging Principle formalises this: any aesthetic expression with low statistical mass in the training data is structurally vulnerable to absorption by the dominant mode through the mathematical mechanics of conditional expectation. Our empirical analysis confirms this is not merely a theoretical concern: DALL-E 3 and Imagen 4 generate outputs that collapse to approximately 10 effective dimensions out of 768, suppress minority cultural diversity by 15% relative to majority cultures, and converge to cosine similarity 0.81 across independently developed models. Crucially, the ablation study demonstrates that this homogenisation is primarily structural: 91% of the majority/minority diversity gap is present in a base model with no preference training, and classifier-free guidance monotonically amplifies the effect. The common attribution of AI aesthetic homogeneity to RLHF or biased curation, while not incorrect, is substantially incomplete.

The cultural implications are significant. Our framework provides technical substance to the critical observation that generative AI enforces a statistical monoculture—not through deliberate ideological design, but through the logic of optimisation itself. The mimicry gap between AI outputs and authentic cultural expression is mathematically proportional to the marginalisation of that culture in the training data. Platform realism is not a metaphor but a mathematical prediction—one now supported by empirical measurement.

If artificial intelligence is to expand rather than flatten our cultural horizons, its fundamental design must be reconceived. Diversity cannot be an afterthought or a data-side fix; it must be encoded in the generative objective itself. We hope this paper contributes a precise technical language, framework, and empirical methodology for that urgent project.

Acknowledgements

[To be added.]

References

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. *arXiv preprint arXiv:2307.01850*, 2023.

Louise Amoore. Machine learning political orders. *Review of International Studies*, 2023.

Zahra Bayramli, Ayhan Suleymanzade, Na Min An, Huzama Ahmad, Eunsu Kim, Junyeong Park, James Thorne, and Alice Oh. Diffusion models through a global lens: Are they culturally inclusive? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31137–31155, Vienna, Austria, July 2025. Association for Computational Linguistics. Also available as arXiv:2502.08914.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhesikan, Aditya Ramesh, et al. Improving image generation with better captions. *OpenAI Technical Report*, 2023.

Homi K. Bhabha. Of mimicry and man: The ambivalence of colonial discourse. *October*, 28: 125–133, 1984.

Homi K. Bhabha. *The Location of Culture*. Routledge, 1994.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1479–1493, 2023.

Meredith Broussard. *More Than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*. MIT Press, 2023.

Allen Chang, Matthew C. Fontaine, Serena Booth, Maja J. Matarić, and Stefanos Nikolaidis. Quality-diversity generative sampling for learning with synthetic data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19805–19812, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems 34*, 2021.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

Philip Good. Permutation tests: A practical guide to resampling methods for testing hypotheses. *Springer Series in Statistics*, 2000.

Google DeepMind. Imagen 4. Google AI Blog, 2025. Accessed via Google Gemini API.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33*, pages 6840–6851, 2020.

Edward Kang. Ground truth tracings (GTT): On the epistemic limits of machine learning. *Big Data & Society*, 2023.

Simon Lindgren. *Critical Theory of AI*. Polity Press, 2024.

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. In *Advances in Neural Information Processing Systems 36*, 2023.

Roland Meyer. Platform realism: AI image synthesis and the rise of generic visual content. *Transbordeur: Photographie Histoire Société*, (9), 2025.

Fabian Offert and Ranjodh Singh Dhaliwal. The method of critical AI studies, a propaedeutic. *arXiv*, 2025.

Fabian Offert and Thao Phan. A sign that spells: Machinic concepts and the racial politics of generative AI. *Journal of Digital Social Research*, 6(4), 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36*, 2023.

Rita Raley and Jennifer Rhee. Critical AI: A field in formation. *American Literature*, 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

Jonathan Roberge and Michael Castelle. Toward an end-to-end sociology of 21st-century machine learning. In *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies*. Springer International Publishing, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Jathan Sadowski. Habsburg AI. Twitter/X post, 13 February 2023. Originally coined on the Machine Kills podcast, 2023. URL `https://x.com/jathansadowski/status/1625245803211272194`.

Jathan Sadowski. Machine's eye view: Postmodern data science and the politics of ground truth. *Science, Technology, & Human Values*, 2025.

Rylan Schaeffer, Joshua Kazdan, Alvan Caleb Arulandu, and Sanmi Koyejo. Position: Model collapse does not mean what you think. *arXiv preprint arXiv:2503.03150*, 2025.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

Liv Skeete. Habsburg AI: When generative models forget what's real. Medium, May 2025. URL `https://medium.com/@livskeete/habsburg-ai-when-generative-models-forget-whats-real`.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems 32*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. Outstanding Paper Award.

Kaitlyn Tiffany. Maybe you missed it, but the internet 'died' five years ago. The Atlantic, August 2021. URL `https://www.theatlantic.com/technology/archive/2021/08/dead-internet-theory-wrong-but-feels-true/619937/`.