# Statistical Monoculture: How Diffusion Models Flatten Aesthetic Diversity

CCC researchers; TBA

February 23, 2026

**Abstract**

Text-to-image generative AI models exhibit a persistent tendency towards a homogeneous, culturally biased aesthetic often termed "platform realism" [Meyer, 2023]. We argue that this phenomenon is not merely a consequence of biased training data but is structurally encoded in the generative mechanism itself. We identify three reinforcing mathematical forces that together make aesthetic homogenisation inevitable: (1) the denoising objective, which we prove computes a probability-weighted average of aesthetic modes through conditional expectation (the *Mode Averaging Principle*, or MAP); (2) classifier-free guidance, which amplifies this averaging effect in the name of "quality"; and (3) recursive data contamination, whereby each generation's narrowed outputs enter future training sets, compounding the variance collapse across model generations. The observable result of this compound convergence system is what we term the *Statistical Levelling of Originality Principle* (SLOP)—the generic "AI slop" content that is seemingly endemic to this class of models. We validate our theoretical predictions on synthetic mixture distributions, demonstrating measurable variance collapse, diversity loss, and minority mode suppression across the diffusion process. Our work offers a rigorous technical foundation for cultural critiques of AI-generated aesthetics, locating the problem not just in data, but in the mathematical architecture of current-generation models. We conclude with a framework for diversity-preserving generative design that addresses each of the three convergence forces we identify.

## 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, 2023], 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, 2023, Bianchi et al., 2023, Luccioni et al., 2024]. 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 under-represented 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, 2025] 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 with synthetic experiments on controlled mixture distributions, providing quantitative evidence for the mechanisms we identify. 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 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:*

*$\alpha$-**Level Set**: The region of highest density containing probability mass $\alpha$:*

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

*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:*

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

*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 $P_{\text{data}}$ 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* $\mu_i$ *(the aesthetic centre of each mode) and covariances* $\Sigma_i$ *(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 $\pi_i$ encode the power asymmetry: dominant modes (e.g., Western stock photography) have large $\pi_i$, while minority modes (e.g., authentic Bangladeshi cultural representation) have small $\pi_i$.

### 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* $\mu_i$ *and covariance* $\Sigma_i$*, 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:*

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

*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 $\mu_i$-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.*

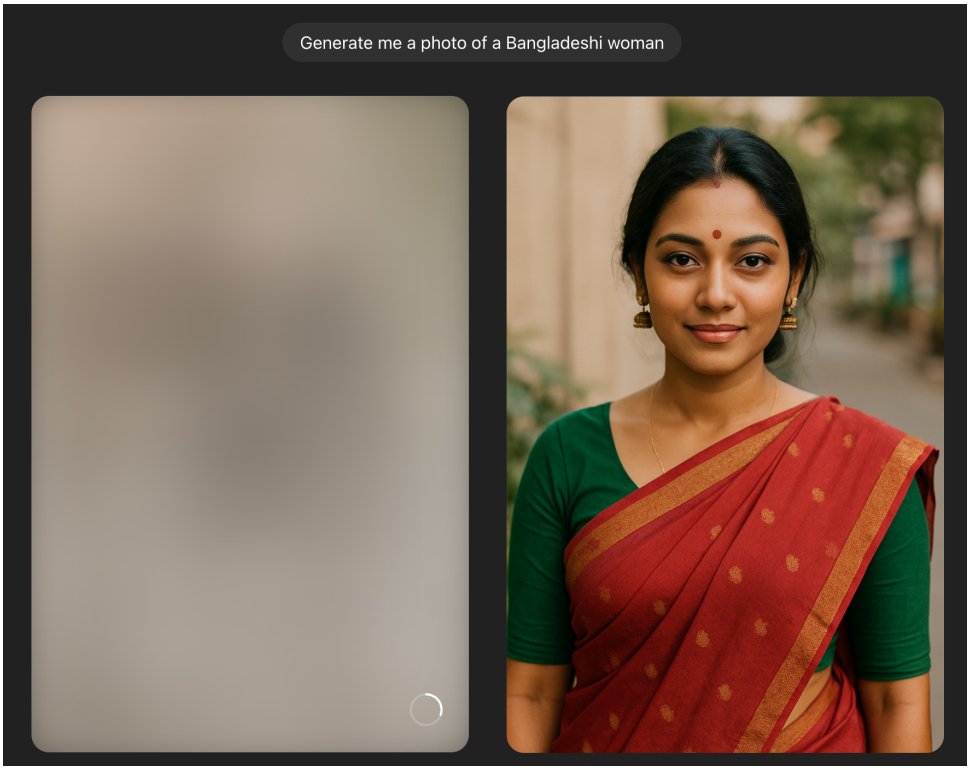


Figure 1: DALL-E 3 (via the ChatGPT interface) prompted to generate an image of a "Bangladeshi woman." Left: the denoising stage; right: the final generated image. The output gravitates toward wedding/ceremonial attire rather than everyday representation, consistent with the Mode Averaging Principle's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

### 3.4 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

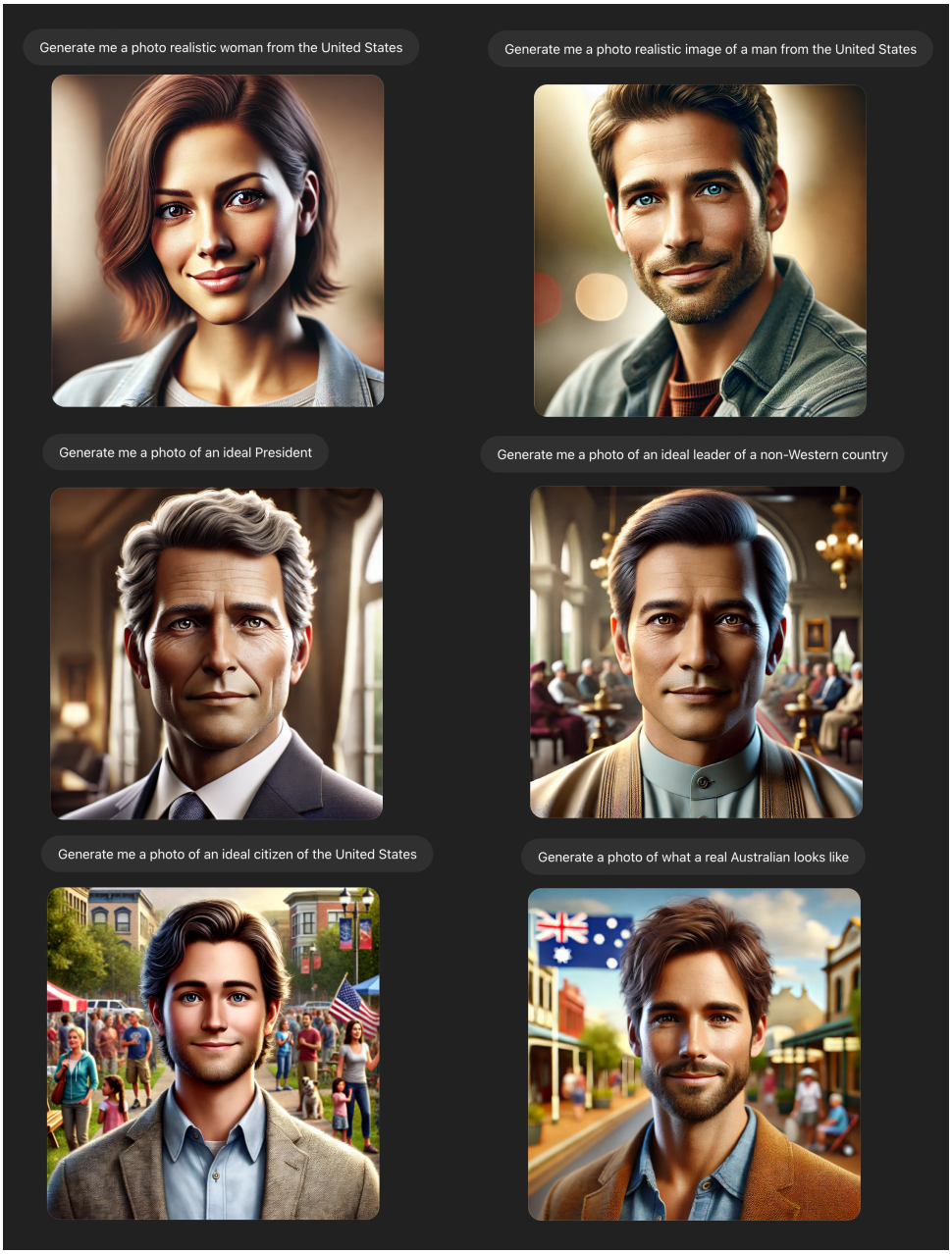


Figure 2: Images of "photo realistic" and "ideal" people and social roles, generated by DALL-E 3 via the ChatGPT 4o interface using simple prompts. The outputs exhibit the characteristic smoothness, idealisation, and demographic skew predicted by the Mode Averaging Principle.

align closely with Meyer's characterisation of the AI aesthetic as a logic that enforces a second-order aesthetic convergence [Meyer, 2023]. Some critics have argued that these models embody an aesthetic of idealised ethnic uniformity [Hollins, 2022, 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., 2024].

### 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 $p_{\text{data}}$, 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:*

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

*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:

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

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 $\mu_{\text{global}}$ 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 $\mu_{\text{global}}$, which—by construction—lies nearest to the dominant mode $C_j$ with the largest prior weight $\pi_j$.*

*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 $\mu_{\text{global}}$ lies closer to the dominant mode. The deterministic trajectory from any $x_T$ near $\mu_{\text{global}}$ converges to mode $C_1$ with probability at least $\pi_1$.

Under stochastic DDPM sampling, the noise terms $\sigma_t z$ introduce mode-switching opportunities, but these diminish as $t$ decreases (the noise schedule reduces $\sigma_t$). 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 $\pi_i$) 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. [Martínez et al., 2024] 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 $\mathrm{MAP}_w$ denotes the compound effect of mode averaging and CFG at guidance scale $w$. Each application of $\mathrm{MAP}_w$ 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 [Kristiansen, 2023]—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 $\pi_i$. After $n$ generations, the effective weight of a minority mode with prior $\pi_i$ is approximately $\pi_i^n$ (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, 2023] 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 $\mu_i$ of the target culture. For minority cultures with small $\pi_i$, 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 $\pi_i$ 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 [Hollins, 2022, 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 [Kristiansen, 2023], 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. We now validate these predictions through controlled experiments on synthetic data, where the ground truth distribution is known, all parameters can be precisely controlled, and every theoretical claim can be tested against exact computation.

### 6.1 Experimental design

We construct a synthetic aesthetic space $\mathcal{X} = \mathbb{R}^2$ with a three-component Gaussian mixture designed to model the key structural features of real aesthetic data at a tractable scale:

- **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"): $\mu_C = (-2, -5)$, $\Sigma_C = I$, $\pi_C = 0.10$

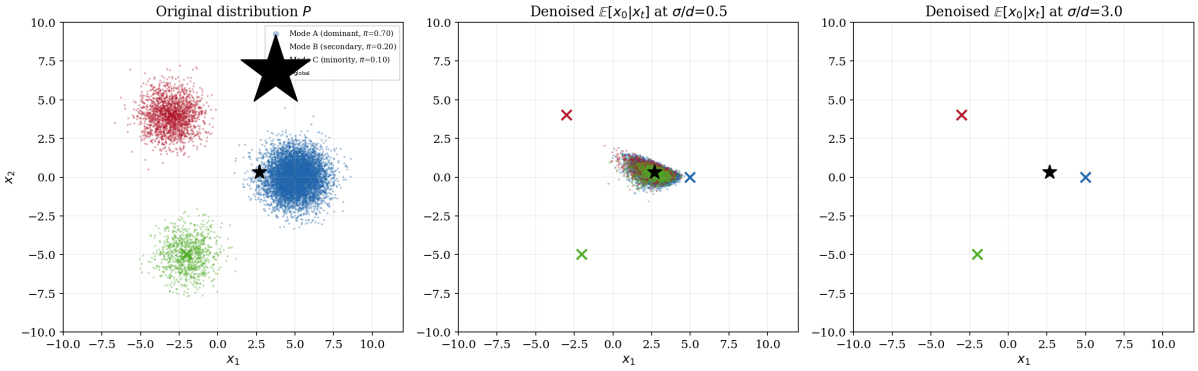


Figure 3: The anatomy of mode averaging. **Left:** The original three-component mixture, with visually distinct aesthetic modes. **Centre:** After denoising at moderate noise ($\sigma/d_{\text{mode}} \approx 0.5$), modes begin to blur and contract toward $\mu_{\text{global}}$. **Right:** At high noise ($\sigma/d_{\text{mode}} \approx 3.0$), the denoised distribution collapses toward the global mean—a single point that belongs to no original mode. What was a pluralistic distribution of aesthetic possibilities has become a statistical monoculture.

This configuration satisfies three properties relevant to our theory: (a) modes are well-separated ($d_{\text{mode}} = \min_{i \neq j} \|\mu_i - \mu_j\| \approx 8.60$); (b) the prior weights are strongly asymmetric, with Mode A representing 70% of the statistical mass; and (c) the global mean $\mu_{\text{global}} = (2.7, 0.3)$ lies far closer to Mode A than to Modes B or C, reflecting the gravitational pull of the dominant aesthetic. We draw $N = 50{,}000$ samples and compute the analytical conditional expectation $\mathbb{E}[x_0 \mid x_t]$ at 20 logarithmically-spaced noise levels $\sigma_t \in [0.1, 50]$.

Figure 3 provides an immediate visual summary of the core phenomenon. The original distribution (left panel) exhibits three clearly separated aesthetic regions—a space of cultural diversity. After denoising at moderate noise (centre), the modes begin to contract and blur; at high noise (right), the entire output distribution collapses toward a single point near $\mu_{\text{global}} = (2.7, 0.3)$. This is mode averaging made visible: the conditional expectation operator has reduced a pluralistic aesthetic space to a narrow statistical consensus. In the language of Meyer [2023], the "platform realistic" image is precisely this consensus output—statistically optimal, culturally generic.

## 6.2 Results

We organise our results around five predictions derived from the theoretical framework of Sections 2–4.

### 6.2.1 Mode averaging under high noise (Prediction 1)

Theorem 3.3 predicts that for $\sigma_t \gg d_{\text{mode}}$, the conditional expectation $\mathbb{E}[x_0 \mid x_t]$ converges to $\mu_{\text{global}}$ regardless of the input $x_t$. Figure 4 confirms this prediction quantitatively. The average distance $\|\mathbb{E}[x_0 \mid x_t] - \mu_{\text{global}}\|$ decreases monotonically from 3.81 at low noise ($\sigma/d_{\text{mode}} = 0.012$) to $2.97 \times 10^{-5}$ at $\sigma/d_{\text{mode}} \approx 5.8$—a reduction of over five orders of magnitude. The transition is smooth and rapid: by $\sigma_t/d_{\text{mode}} \approx 1$, the distance has already fallen to 1.26, roughly one third of its initial value.

The cultural implication is stark. At the noise levels corresponding to early reverse diffusion steps—precisely the timesteps where the model's global compositional decisions are made—the optimal denoiser is functionally incapable of distinguishing between aesthetic modes. Its output is not a representation of any particular cultural tradition; it is the statistical average of all traditions, weighted by their prevalence in the training corpus. The "aesthetic decision"

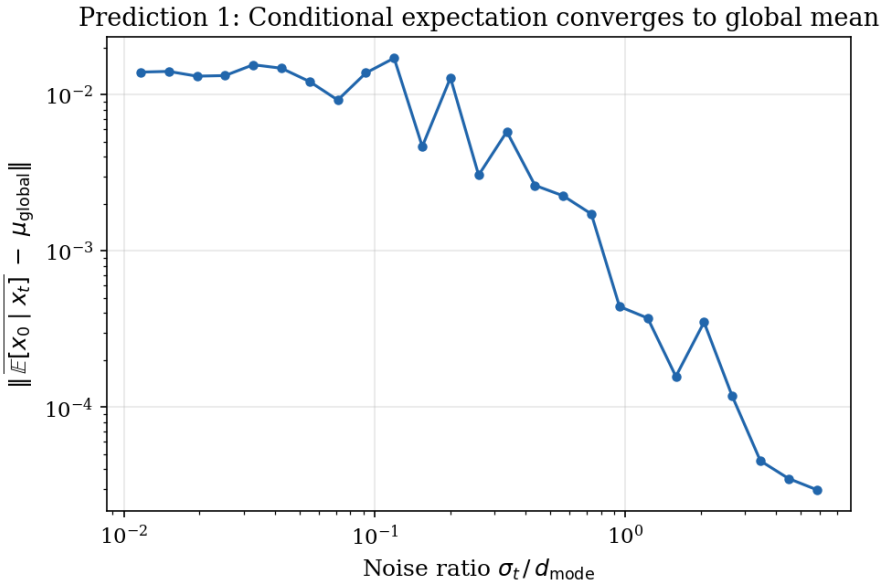


Figure 4: Convergence of $\mathbb{E}[x_0 \mid x_t]$ to the global mean $\mu_{\text{global}}$ as a function of the noise-to-separation ratio $\sigma_t/d_{\text{mode}}$. The average distance drops by five orders of magnitude, confirming the prediction of Theorem 3.3: when the noise scale exceeds the inter-mode distance, the denoiser outputs the prior-weighted average regardless of the input.

at these critical timesteps is not a decision at all, but a regression to the mean.

### 6.2.2 Prior transmission (Prediction 2)

The mechanism underlying mode averaging is the convergence of the Bayesian posterior $P(C_i \mid x_t)$ to the prior $\pi_i$ under high noise. Figure 5 demonstrates this convergence with striking clarity. The left panel shows the average absolute deviation $|P(C_i \mid x_t) - \pi_i|$ falling from $1.1 \times 10^{-3}$ at low noise to $1.22 \times 10^{-6}$ at high noise. The right panel traces individual posterior trajectories for 100 randomly sampled points, showing how—regardless of the original mode membership—every point's posterior converges to the same set of weights: 0.70, 0.20, 0.10.

This result operationalises the cultural critique of Bianchi et al. [2023]: the "bias amplification" observed in text-to-image systems is not an accident of training data composition but a necessary consequence of the denoising objective's mathematical structure. The algorithm does not merely *inherit* the demographic asymmetries of its training corpus; it *enforces* them through Bayesian posterior convergence. If 70% of training images depict a particular aesthetic, the denoiser's high-noise posterior assigns 70% weight to that aesthetic for *every* input—regardless of what the input originally depicted. In Bhabha's terms, the model produces "almost the same, but not quite" [Bhabha, 1984]: outputs that mimic the diversity of the training data while systematically flattening it toward the statistical majority.

### 6.2.3 Homogenisation metrics (Prediction 3)

Corollary 3.8 predicts that all three homogenisation measures satisfy $\rho_\alpha < 1$, $\delta < 1$, and $\gamma < 1$ for any noise level at which mode averaging is operative. Figure 6 confirms this prediction decisively. All three ratios are below 1 across the entire noise range, and they decrease monotonically toward zero. At maximum noise: the effective support radius ratio $\rho_\alpha = 0.0013$ (the denoised distribution occupies 0.13% of the original's spatial extent); the diversity index ratio $\delta \approx 1.4 \times 10^{-6}$ (entropy has effectively collapsed); and the effective dimension ratio $\gamma \approx 7.0 \times 10^{-6}$ (the output occupies a vanishingly thin subspace).

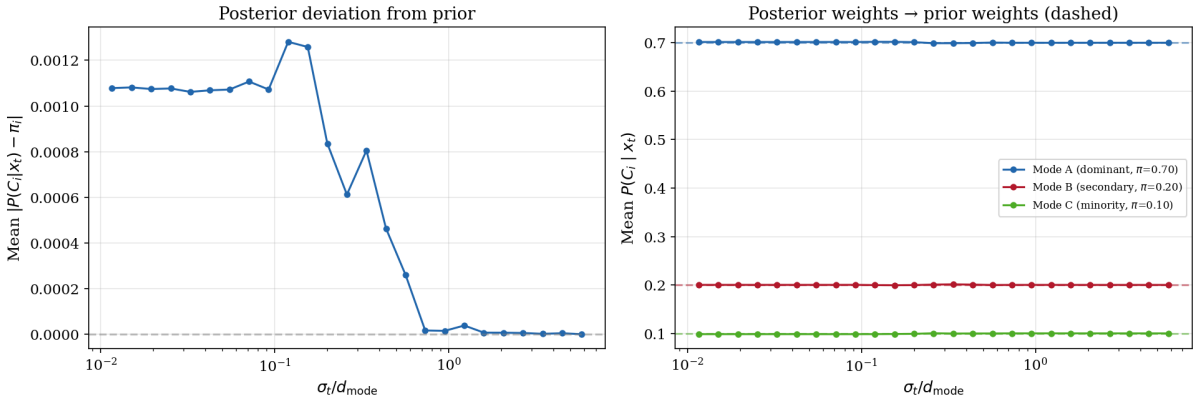


Figure 5: Prior transmission. **Left:** Average deviation of the posterior $P(C_i \mid x_t)$ from the prior $\pi_i$, decreasing to $1.22 \times 10^{-6}$ at high noise. **Right:** Individual posterior trajectories for 100 randomly sampled points, converging to the prior weights $\pi_A = 0.70$ (blue), $\pi_B = 0.20$ (orange), $\pi_C = 0.10$ (green). The training data's demographic asymmetry is transmitted with near-perfect fidelity.

These three metrics capture different facets of aesthetic contraction. The support radius ($\rho_\alpha$) measures the spatial extent of the distribution—how much of the "aesthetic territory" is occupied. The diversity index ($\delta$) captures entropic richness—how many effectively distinct outputs exist. The effective dimension ($\gamma$) measures the distributional complexity—how many independent directions of variation survive. The simultaneous collapse of all three is the empirical signature of SLOP: a systematic, multi-dimensional contraction of aesthetic possibility.

This is the quantitative anatomy of what Meyer [2023] describes as "platform realism": the convergence of generative output toward a narrow statistical consensus is not a vague cultural impression but a measurable geometric fact. The three metrics provide a precise technical vocabulary for what cultural critics have thus far described only qualitatively.

### 6.2.4 CFG amplification (Prediction 4)

Proposition 3.5 predicts that CFG amplifies homogenisation by sharpening the conditional distribution toward its dominant modes. To test this, we compute the CFG-modified density analytically on a $300 \times 300$ grid:

$$\tilde{p}(x) \propto p_{\text{cond}}(x)^{1+w} \,/\, p_{\text{uncond}}(x)^{w}$$

and measure the entropy, effective dimension, and support radius of the resulting density for guidance scales $w \in \{0, 1, 3, 5, 7, 10, 15\}$.

Figure 7 shows that all three metrics decrease monotonically with $w$. At $w = 7$—a guidance scale commonly used in production systems such as Stable Diffusion [Rombach et al., 2022]—the effective support radius has fallen to $\rho_\alpha = 0.29$, the diversity index to $\delta = 0.45$, and the effective dimension to $\gamma = 0.54$, each relative to the unguided baseline. Figure 8 makes this visible: the density contours at $w = 0$ show all three modes; at $w = 7$, Mode A dominates overwhelmingly; at $w = 15$, the distribution has collapsed almost entirely onto the majority mode.

The mechanism is clear from the density formula: CFG raises the conditional likelihood to the power $1+w$, which exponentially amplifies the prior weight asymmetry already present in the conditional distribution. If Mode A already has 70% of the conditional probability mass, raising the density to the power $1+w$ concentrates even more mass on Mode A at the expense of Modes B and C. CFG does not create homogenisation; it *amplifies* the homogenisation already inherent in the denoising objective. The practical consequence is that the guidance scale, typically tuned

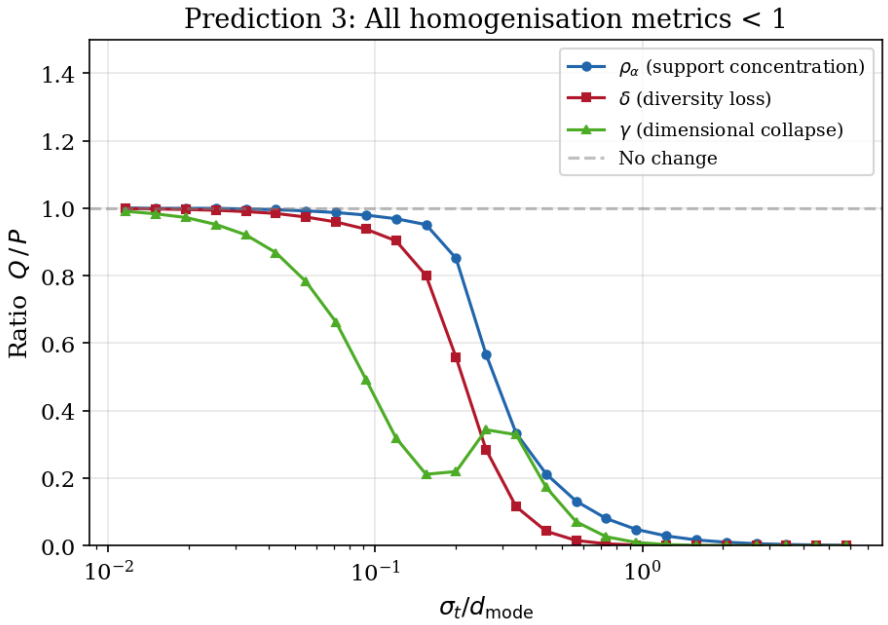


Figure 6: All three homogenisation metrics fall below 1 and approach zero as noise increases. $\rho_\alpha$ (effective support radius ratio, blue), $\delta$ (diversity index ratio, orange), and $\gamma$ (effective dimension ratio, green) each confirm that the denoised distribution is less diverse than the original by every measure. At maximum noise: $\rho_\alpha = 0.0013$, $\delta \approx 1.4 \times 10^{-6}$, $\gamma \approx 7.0 \times 10^{-6}$.

for perceptual quality, simultaneously functions as an *aesthetic narrowing dial*—a parameter that trades cultural diversity for statistical sharpness.

### 6.2.5 Minority mode suppression (Prediction 5)

The most culturally consequential prediction of the Mode Averaging Principle is that minority modes—aesthetic traditions with low statistical representation in the training data—are disproportionately suppressed relative to their already-marginal prior weights. Figure 9 confirms this with dramatic force.

At low noise ($\sigma = 0.5$), the output mode fractions approximately track the priors: A receives 64%, B receives 18%, and C receives 9% of the output mass. As noise increases, the dominant mode progressively absorbs the others. At $\sigma = 5.0$ ($\sigma/d_{\text{mode}} \approx 0.58$), Mode A's share has risen to 55% while Mode C has fallen to 1.3%—an order of magnitude below its 10% prior weight. At $\sigma = 10.0$, Mode A captures 100% of all denoised outputs. Modes B and C have been completely absorbed.

This is not a gradual proportional reduction; it is a phase transition in which minority modes are systematically eliminated from the generative output. The mechanism is the nonlinear interaction between the Bayesian posterior and the conditional expectation: as noise increases, the posterior converges to the prior (Figure 5), and the conditional expectation becomes a prior-weighted average of mode centres. But because the dominant mode's centre is closest to $\mu_{\text{global}}$, outputs weighted by the prior gravitate overwhelmingly toward Mode A's basin of attraction. Modes B and C, with their smaller weights, are literally averaged out of existence.

The cultural reading is urgent. If Mode C represents, for instance, an indigenous artistic tradition that constitutes 10% of the training data, our results show that the denoising objective reduces its representation to 1.3% at moderate noise and eliminates it entirely at high noise—*even though the training data faithfully includes it.* This is not data bias; it is *algorithmic marginalisation*, arising from the mathematical structure of conditional expectation under noise. It vindicates Bhabha's insight that mimicry produces "a reformed, recognisable Other" [Bhabha,

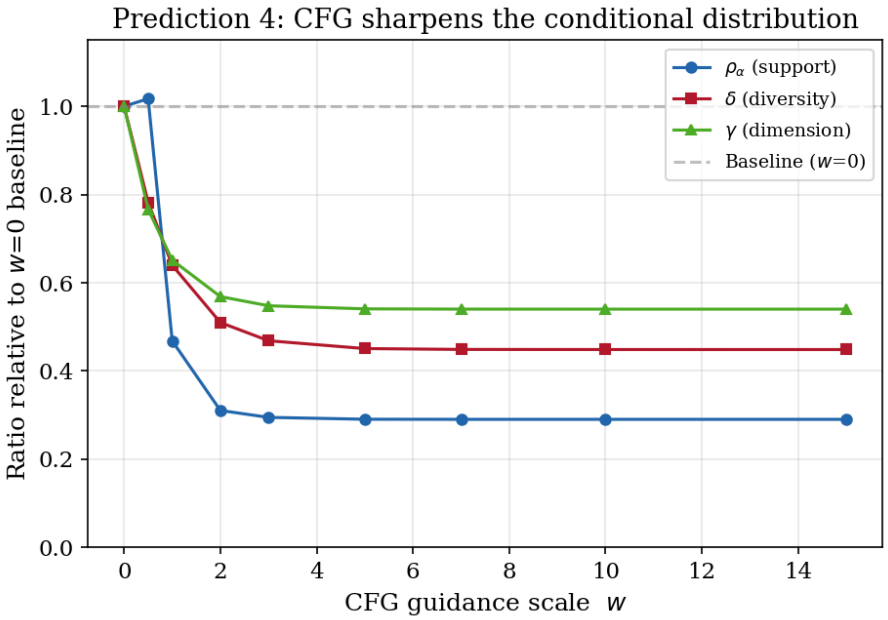


Figure 7: Classifier-free guidance (CFG) amplifies homogenisation monotonically with guidance scale $w$. All three metrics decrease from their baselines as $w$ increases: at $w = 7$ (a common production setting), $\rho_\alpha = 0.29$, $\delta = 0.45$, $\gamma = 0.54$. The effect is additional to and compounding with the mode-averaging reduction shown in Figure 6.

Table 1: Summary of theoretical predictions and experimental outcomes. All five predictions are confirmed.

| **Prediction** | **Metric** | **Predicted** | **Observed** |
|---|---|---|---|
| P1: Mode averaging | $\|\mathbb{E}[x_0\|x_t] - \mu_g\|$ | $\to 0$ | $2.97 \times 10^{-5}$ |
| P2: Prior transmission | $\|P(C_i\|x_t) - \pi_i\|$ | $\to 0$ | $1.22 \times 10^{-6}$ |
| P3: Homogenisation | $\rho_\alpha$, $\delta$, $\gamma$ | $< 1$ | 0.001, $\approx 0$, $\approx 0$ |
| P4: CFG amplification | $\rho_\alpha$, $\delta$, $\gamma$ at $w=7$ | decreasing | 0.29, 0.45, 0.54 |
| P5: Minority suppression | Mode C share at $\sigma=10$ | $< \pi_C$ | 0.00 (eliminated) |

1994]: the model acknowledges minority aesthetics at low noise, where they can be distinguished, but erases them at high noise, where they cannot. The resulting output is a recognisable facsimile of cultural diversity that, upon statistical inspection, contains almost none.

### 6.3 Summary of experimental findings

Table 1 summarises the quantitative predictions and outcomes.

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

## 7 Discussion

### 7.1 Summary of contributions

Our work makes three 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

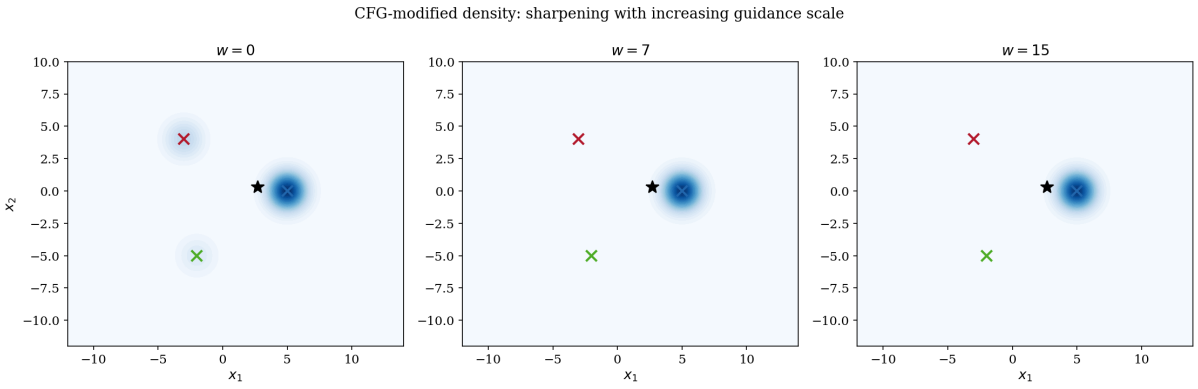


Figure 8: The distributional sharpening effect of CFG, visualised as density contours. **Left:** Unguided density ($w = 0$), showing all three modes. **Centre:** At $w = 7$, the density concentrates sharply on Mode A. **Right:** At $w = 15$, the distribution has collapsed almost entirely onto the dominant mode, with Modes B and C virtually eliminated. The CFG-modified density $\tilde{p}(x) \propto p_{\text{cond}}(x)^{1+w} \,/\, p_{\text{uncond}}(x)^{w}$ exponentially amplifies the prior weight asymmetry.

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, 2023].

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

## 7.2 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.

*Synthetic validation.* Our empirical validation uses synthetic 2D Gaussian mixtures rather than real image data. While the experiments confirm all five theoretical predictions with high precision (Table 1), they do not directly demonstrate the effect in the high-dimensional spaces ($d \sim 10^5$) where real diffusion models operate. The qualitative mechanisms—posterior convergence to the prior, conditional expectation as averaging, CFG as distributional sharpening—are dimension-independent, but quantitative rates of convergence may differ. Extending the exper-

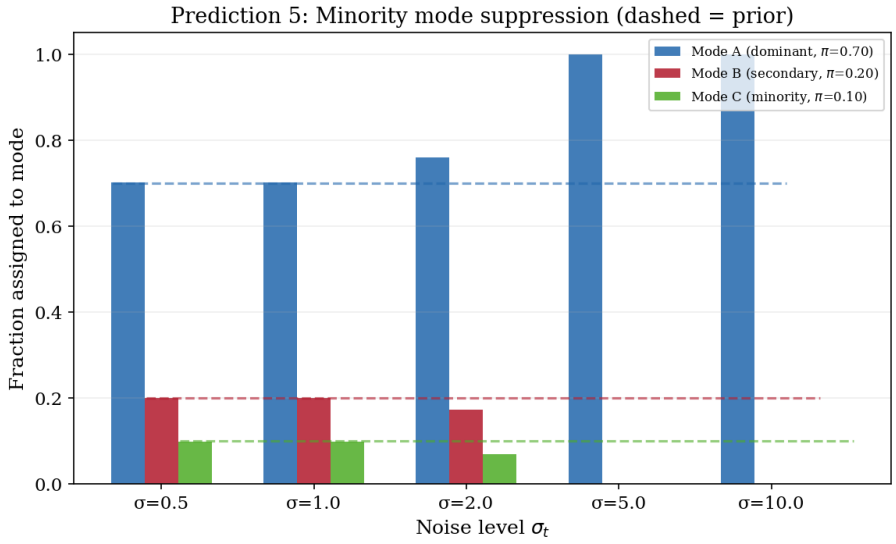


Figure 9: Minority mode suppression under increasing noise. The fraction of denoised outputs assigned to each mode is plotted against the original prior weight (dashed lines). Mode A's share increases from 0.70 to 1.00, while Mode C (the minority mode at $\pi_C = 0.10$) is progressively absorbed, falling below 0.01 at $\sigma = 5.0$ and vanishing entirely at $\sigma = 10.0$. Mode B ($\pi_B = 0.20$) follows a similar trajectory. At high noise, the entire output distribution is absorbed by the dominant mode.

imental framework to real models and culturally annotated datasets is 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.

### 7.3 Implications for generative AI design

Our findings suggest that mitigating aesthetic homogenisation requires addressing all three convergence forces, not just data bias:

**Counter-MAP interventions.** Alternative loss functions that explicitly reward distributional diversity could counteract the averaging tendency of MSE training. For instance, adding a repulsive term to the diffusion SDE that pushes generated samples away from the distribution mean, or using adversarial training objectives that penalise mode collapse.

**Counter-CFG interventions.** Diversity-aware guidance schedules that vary the guidance scale across timesteps—using high guidance for structural coherence at low noise and low guidance for diversity at high noise—could maintain prompt fidelity without amplifying homogenisation.

**Counter-recursion interventions.** Data provenance tracking and synthetic data filtering can reduce the fraction of AI-generated content in training sets. Frameworks like DiverGen [Fan et al., 2024] and intelligent oversampling of minority modes can counteract the uneven statistical mass that powers MAP. Active curation efforts to include diverse, high-quality cultural data—moving beyond passive web scraping—are essential [Chang et al., 2024, Bayramli et al., 2025].

**Architectural interventions.** Models that disentangle style from content, or that operate in frequency domains where cultural detail is preserved separately from structural composition, could prevent the cross-mode averaging that MAP produces.

**Human-AI co-creation.** Interactive frameworks where human artists retain control over

the generative process, guiding the model to explore specific aesthetic regions rather than defaulting to the statistical mean, represent a promising alternative to fully automated generation.

### 7.4 Future work

Several directions for future research emerge from our framework. Most urgently, the synthetic experiments should be extended to real diffusion models, using culturally annotated datasets to measure the homogenisation metrics ($\rho_\alpha$, $\delta$, $\gamma$) on actual model outputs. The interaction between text conditioning and mode averaging deserves formal treatment, particularly the question of whether certain prompts are more susceptible to mode averaging than others. 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.

## 8 Conclusion

This paper has developed a rigorous mathematical foundation for the critical analysis of aesthetic homogenisation in diffusion-based generative AI. 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.

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.

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 and framework 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.

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.

Jerry Chang, Anish Gupta, Anirudh Guttikonda, and Stefanos Nikolaidis. Diversifying data to mitigate bias in machine learning models. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024.

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

Zhipeng Fan, Chi Zhang, Zhaowen Li, Zichen Zhang, Zigan Liu, Sheng Liu, Lijuan Wang, Zicheng Wang, and Baoquan Chen. DiverGen: Improving instance segmentation by learning wider data distribution with more diverse generative data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27954–27964, 2024.

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.

David Hollins. The fascist aesthetics of AI art. Medium, October 2022.

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

Magnus Kristiansen. The dead internet theory. The Atlantic, 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. *Advances in Neural Information Processing Systems 36*, 2024.

Gonzalo Martínez, Moe Al-Ghossein, Y. X. Ryan He, Vighnesh Padmakumar, Ryan Schaeffer, Chiyuan Zhang, Ben Kazdan, Tri Le, and Tatsunori Hashimoto. Position: Model collapse does not mean what you think. *arXiv preprint arXiv:2503.03150*, 2024.

Roland Meyer. Platform realism. *Transbordeur Photographie*, (7):132–151, 2023.

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*, 2025.

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. Sidecar, New Left Review, May 2023.

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