A Zipfian Analysis of Visual Token Distributions for AI-Generated Images

Andrew Shin
Keio University
Yokohama, Kanagawa Prefecture, Japan
shin@ics.keio.ac.jp

Abstract

The rapid evolution of text-to-image generation has blurred the perceptual boundary between natural and synthetic imagery. However, it remains questionable whether the statistical structure of generated visual content mirrors the information density of the physical visual world. Drawing upon principles from statistical linguistics, this study investigates the visual language of generative models through the lens of Zipfian dynamics. By analyzing a large-scale corpus of real and synthetic images, we uncover a fundamental divergence between visual syntax and semantics. We find that while generative models have successfully replicated the low-level physics of light, their high-level texture vocabulary exhibits distinct statistical signatures. Our analysis reveals a spectrum of entropy, identifying architectural fingerprints unique to each model. Furthermore, we investigate the relation ship between generated images and prompt complexity, and find that increasing the semantic specificity of text prompts systematically degrades the statistical realism of the generated output.

1 Introduction

The field of computer vision has witnessed a paradigm shift with the advent of diffusion-based text-to-image generation. State-of-the-art image generation models have largely surmounted the previous hurdles of synthetic imagery, and the typical artifacts such as malformed hands, incoherent lighting, or surreal geometry are vanishing at a rapid pace. As these systems achieve photorealistic fidelity, the perceptual boundary between natural photography and synthetic media is eroding, rendering traditional, artifact-based forensic methods increasingly obsolete. However, the ability to clone the *appearance* of reality does not necessarily equate to replicating the underlying *stochastic processes* that govern the natural world.

This divergence prompts a fundamental question: does the generative process leave an invisible statistical fingerprint? While a generated image may be indistinguishable from a photograph to the human eye, it is ultimately the product of a mathematical optimization process of denoising trajectory guided by a text encoder, rather than a physical capture of photons interacting with matter. We posit that natural images, like natural language, possess a hierarchical statistical structure governed by universal laws of information density and entropy, as has been suggested by the pioneering works on analyzing statistical behavior of visual data with linguistic laws (Ruderman, 1996; Crosier and Griffin, 2007; Chan et al., 2024; Tsai et al., 2025), which directly inspire our work in this paper. If generative models are fundamentally statistical mimics, they may inadvertently suppress the chaotic, heavy-tailed entropy characteristic of the physical world in favor of cleaner, more probable visual patterns.

In this paper, we use these linguistic principles to audit the visual language of generative AI. We analyze a large-scale corpus of 50,000 images, stratified across real photography and state-of-the-art generative models. We employ a dual-tokenization strategy, separating visual physics (low-level edges) from texture vocabulary (high-level visual words).

Our contributions are threefold. 1) We demonstrate a dissociation between syntax and semantics: while AI models have effectively solved the Zipfian statistics of low-level features, they diverge significantly in their texture vocabulary, following log-normal rather than pure power law dynamics. 2) We also identify a spectrum of entropy that fingerprints model architecture, ranging from the artificially ordered, low-entropy distributions to the real, chaotic statistics. 3) We uncover its relevance to prompt complexity, providing empirical evidence that increasing the semantic specificity of a text prompt acts as an entropy constraint, degrading the statistical realism of the generated output.

2 Related Work

The inquiry into the statistical structure of natural scenes was pioneered by (Ruderman, 1996), who proposed that the scale-invariance of natural images arises because the visual world is a collage of statistically independent objects following a power-law size distribution. Building on this foundation, (Crosier and Griffin, 2007) formalized the analogy between vision and language. They defined an basic image features (BIFs), classifying pixels into geometric primitives like edges and blobs, and treated $n \times n$ patches of these features as visual words. They demonstrated that, for specific parameter settings, these visual words strictly obey Zipf's Law (Zipf, 1932), arguing that this distribution represents an optimally efficient code for object recognition, analogous to the efficiency of natural language.

Recent research has extended these findings from hand-crafted features to the learned representations of deep neural networks. (Tsai et al., 2025) utilized the kernels of pre-trained CNNs such as VGG-19 (Simonyan and Zisserman, 2014) to define visual words based on activation intensity. By analyzing layer-wise statistics, they confirmed that not only Zipf's Law but also Heaps' Law (Heaps, 1978) with vocabulary growth and Benford's Law (Benford, 1938) with leading digit distribution also emerge in the visual words. Their findings suggest a correspondence between the evolution of visual symbols in deep networks and the structure of human language.

With the advent of autoregressive vision transformers, (Chan et al., 2024) provided a comprehensive audit of discrete visual languages, such as tokens in LLaVA (Liu et al., 2023) and Chameleon (Team et al., 2024). While they confirmed that discrete visual tokens follow Zipfian distributions, they identified critical structural differences from NLP: visual languages exhibit higher perplexity, weaker hierarchical grammar, and a tendency for tokens to represent intermediate granularity such as object parts, rather than semantic wholes. Crucially, their work focused on the properties of the internal representation of real images, whereas our work investigates the statistical realism of the generative output produced by diffusion models.

3 Methodology

To compare the statistical properties of real vs. AI imagery, we employ a dual-tokenization strategy, analyzing both low-level features (physics) and high-level textures (semantics).

3.1 Feature Extraction Strategy

Basic Image Features (BIFs): To capture the low-level structure of the image (edges, blobs, flat regions), we compute BIF maps. Following (Crosier and Griffin, 2007), we compute the response of the image I to a bank of Gaussian derivative filters up to the second order at scale σ:

$$s_{nm} = \sigma^{n+m} \cdot \left(\frac{\partial^{n+m}}{\partial x^n \partial y^m} G_\sigma \right) * I \qquad (1)$$

Pixels are classified into one of 7 symmetry classes (flat, slope, dark blob, light blob, dark line, light line, saddle) based on the invariant properties of the local jet. This tokenization represents the "syntax" of the visual world, i.e. the fundamental rules of geometry and contrast.

Vector Quantization (VQ): To capture complex textures (e.g., fur, grass, fabric), we employ vector quantization. We extract non-overlapping 4×4 pixel patches p_i from all images. We learn a shared vocabulary $V = \{v_1, ..., v_K\}$ of size $K = 512$ using mini-batch K-means on a mixed sample of real and synthetic patches to establish a statistically neutral, unified basis for comparison. Each patch is then assigned to its nearest visual word:

$$w_i = \arg\min_k ||p_i - v_k||_2^2 \qquad (2)$$

This creates a discrete "document" of visual words for every image, allowing us to analyze the diversity and repetition of textures.

3.2 Statistical Analysis

We fit the frequency distribution of tokens to a discrete power law distribution:

$$P(x) = Cx^{-\alpha}, \quad x \geq x_{\min} \qquad (3)$$

where α is the scaling parameter and $x_{\min}$ is the lower bound of the power-law behavior. We use the Kolmogorov-Smirnov distance to estimate $x_{\min}$ and maximum likelihood estimation for α. To confirm the validity of the fit, we compare it against a log-normal distribution via a likelihood ratio test ($\mathcal{R}$), where $\mathcal{R} > 0$ supports the power law and $\mathcal{R} < 0$ supports log-normal.

4 Experiments

4.1 Dataset

We utilize the Rapidata Image Generation Alignment Dataset (Rapidata, 2025) to source 40,000 AI-generated images, stratified equally (10,000 each)

Table 1: Aggregate analysis ($N = 50,000$). For BIFs, AI is indistinguishable from reality ($p > 0.05$). For VQ, both distributions are log-normal ($R < 0$), but AI exhibits a steeper slope and lower $x_{\min}$.

Dataset	**Alpha** (α)	$x_{\min}$	$\mathcal{R}$	p**-val**
BIF				
Real (Aggregate)	1.8180	34.9M	0.002	0.70
AI (Aggregate)	1.7608	32.8M	0.003	0.85
VQ				
Real (Aggregate)	1.9490	1.93M	-13.52	< 0.01
AI (Aggregate)	2.2868	1.42M	-6.60	< 0.01

across four state-of-the-art models, namely DALL-E 3 (Betker et al.), Midjourney (Midjourney, 2024), Stable Diffusion 3 (SD3) (Esser et al., 2024), and Flux 1.1 Pro (Labs et al., 2025). For the real-world baseline, we sample 10,000 images from ImageNet-1k (Russakovsky et al., 2015) validation set, representing a general purpose distribution of natural photography. Feature extraction and statistical analysis described in Section 3 are then applied to the collected data.

4.2 Results

4.2.1 Aggregate Analysis

Table 1 presents a fundamental dichotomy in the statistical footprint of generative models: a near-perfect replication of low-level visual syntax, contrasted against a systematic deviation in high-level texture semantics.

The analysis of BIFs reveals that, at the level of local geometry, AI-generated imagery is statistically indistinguishable from natural photography. The high p-values and negligible likelihood ratios suggest that the distribution of fundamental geometric primitives is invariant across real and synthetic domains.

In contrast, the visual vocabulary analysis with VQ uncovers a significant distributional shift. Consistently negative likelihood ratios for both datasets indicate that visual word frequencies follow a log-normal distribution rather than a pure power law. AI aggregate exhibits a steeper decay, which may be attributed to suppression of entropy, where the generative process under-samples the rare, chaotic textures that populate the "long tail" of reality.

Furthermore, the difference in $x_{\min}$, which marks the transition point between head and tail of the distribution, is diagnostic. Real images maintain a significantly higher cutoff, implying an increased density at the upper tail. Nature sustains a larger diversity of common textures before the distribution transitions into its decaying tail behavior. AI models, by contrast, transition to the decay phase earlier, collapsing the diversity of visual textures.

Table 2: Model-specific analysis. All models match real-world images in BIF with $p \approx 0.9$. With VQ, however, models display distinct behaviors.

Model	**Alpha** (α)	$x_{\min}$	$\mathcal{R}$	p**-val**
BIF Analysis (Physics)				
Stable Diffusion 3	1.8109	7.89M	0.001	0.94
DALL-E 3	1.8026	9.32M	0.001	0.95
Flux 1.1 Pro	1.7857	6.91M	0.001	0.89
Midjourney	1.7318	5.62M	0.001	0.93
VQ Analysis (Texture)				
DALL-E 3	2.4388	391k	-13.95	< 0.01
Stable Diffusion 3	2.2909	258k	-6.10	< 0.01
Midjourney	2.2513	379k	-17.45	< 0.01
Flux 1.1 Pro	2.1741	246k	-18.54	< 0.01

4.2.2 Model-Specific Analysis

Breaking down the aggregate results reveals distinct statistical fingerprints for each architecture (Table 2). While all models converge on the physics of low-level features, they diverge significantly in how they manage the entropy of high-level textures, which may be linked to architectural choices.

DALL-E 3 exhibits the steepest Zipfian slope, indicating the least entropic texture distribution. This observation is consistent with DALL-E 3's integration of LLM guidance for prompt rewriting. We hypothesize that minimizing the semantic gap between text and image may inadvertently act to reduce the stochastic variance of the output and suppress the long tail of messy, unprompted textures in favor of high-probability visual concepts.

Stable Diffusion and Midjourney occupy a middle ground, likely driven by distinct optimization constraints. For Stable Diffusion, the reliance on latent diffusion involves compressing data into a lower-dimensional latent space via a VAE. This process may have acted as a low-pass filter, discarding the high-frequency Zipfian tail of pixel noise to ensure stability. Midjourney's similarity may have stemmed from aggressive RLHF tuned for aesthetics, which biases the model toward clean and statistically regular textures, while penalizing the chaotic aspects of the physical world.

Flux exhibits the lowest slope among the AI models, placing it statistically closest to the high-entropy nature of real-world data. This could potentially be attributed to its *flow matching* paradigm. Unlike standard diffusion, which iteratively re-

Table 3: Impact of prompt complexity. Specificity increases the Zipfian slope, reducing statistical realism.

Prompt Type	**Alpha** (α)	$x_{\min}$
Simple ($<$ 6 words)	1.9283	132,245
Complex ($>$ 15 words)	2.3259	197,883

moves noise and may smooth out high-frequency irregularities, flow matching models optimize transport paths, which may allow the model to preserve a higher degree of high-frequency variance, resulting in a texture distribution that feels more physically plausible.

While no model perfectly replicates the heavy-tailed entropy of natural photography, the data suggests a trade-off; stronger semantic guidance appears to correlate with reduced statistical realism, whereas newer flow-based architectures seem better equipped to sustain the chaotic vocabulary of the visual world.

5 The Effect of Prompt Complexity

Images generated by text-to-image generation models are naturally influenced by the input prompts. A prevalent assumption in prompt engineering is that increasing the lexical density of a prompt enhances the realism of the generated output. We investigate this hypothesis by analyzing whether conditioning on complex, lengthy prompts increases the statistical diversity of the generated visual vocabulary, or whether it inadvertently constrains the model's latent space to a narrower manifold of learned visual concepts.

5.1 Experimental Setting and Procedure

We stratified the data into two distinct groups based on prompt word count:

- **Simple Prompts ($<$ 6 words):** These prompts typically consist of a single subject or short phrase (e.g., "A red car," "A cat"). They provide minimal semantic constraints, requiring the model to "hallucinate" the majority of the scene's details, such as background, lighting, style, from its unconditioned prior.
- **Complex Prompts ($>$ 15 words):** These prompts are highly descriptive, often specifying multiple attributes such as lighting conditions, artistic style, texture quality, and background elements (e.g., "A futuristic city skyline at sunset with neon lights reflecting on wet pavement, 4k, photorealistic").

We collected a balanced sample of $N = 5,000$ images per group. To ensure a unified basis for comparison, we trained a shared VQ codebook on the combined dataset. We then tokenized all images into visual word sequences and performed maximum likelihood estimation to determine the Zipfian parameters for each group.

5.2 Results and Analysis

Table 3 presents the results of the distributional analysis.

The Zipfian slope for complex prompts is steeper than for simple prompts, indicating a faster decay in the frequency of rare events. This suggests that when the model is heavily conditioned by a long text string, it relies more heavily on a core set of high-probability visual tokens to satisfy the semantic constraints, suppressing the generation of rare, stochastic textures that characterize the heavy tail of natural imagery. On the other hand, conditioning on simple prompts allows the model to sample more freely from its diverse training distribution. In addition, the complex condition exhibits a significantly higher $x_{\min}$, implying that the distribution for complex prompts is dominated by a larger block of repetitive, common visual words before the Zipfian tail behavior emerges.

In summary, contrary to the intuition that more text leads to more visual information, we observe a systematic degradation in statistical realism as prompt complexity increases, and specificity in language appears to constrain the entropy of the visual generation process.

6 Conclusion

In this study, we examined the statistical behavior of AI-generated images with Zipfian approach. Our results demonstrate that, while generative AI has effectively mastered the low-level physics, it exhibits a distinct statistical fingerprint in its high-level texture vocabulary, whether in aggregate or with distinct models. We also observe that architectural choices of each model may be linked to their statistical behaviors. Finally, we examined the the effect of prompt complexity in the statistical behavior of generated images, and find that, contrary to the popular intuition, increasing prompt complexity acts as an entropy constraint, driving the output distribution further from the heavy-tailed diversity of the physical world.

Limitations

Our study faces several methodological constraints. First, our visual word tokenization relies on shallow K-Means clustering of raw pixel patches. While this heuristic aligns with classical vision literature, it captures local texture statistics rather than the high-level semantic tokens utilized by modern latent transformers. Consequently, our findings reflect the entropy of surface-level texture rather than deep semantic structure. Additionally, our analysis was conducted at a fixed scale ($\sigma = 1.0$), potentially overlooking scale-invariant statistical divergences that may exist at macro-compositional levels or sub-pixel resolutions.

Second, our dataset choice introduces a potential selection bias, as the images were generated to maximize aesthetic alignment rather than representing the raw, rejection-free output of the models. Furthermore, our investigation of the prompt complexity relied solely on word count as a proxy for semantic density.

References

Frank Benford. 1938. The law of anomalous numbers. *Proceedings of the American Mathematical Society*.

James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. Improving image generation with better captions.

David M. Chan, Rodolfo Corona, Joonyong Park, Cheol Jun Cho, Yutong Bai, and Trevor Darrell. 2024. Analyzing the language of visual tokens. *ArXiv*, abs/2411.05001.

Michael Crosier and Lewis D. Griffin. 2007. Zipf's law in image coding schemes. In *British Machine Vision Conference*.

Patrick Esser, Sumith Kulal, A. Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *ArXiv*, abs/2403.03206.

H. S. Heaps. 1978. Information retrieval, computational and theoretical aspects.

Black Forest Labs, Stephen Batifol, A. Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Muller, Dustin Podell, Robin Rombach, Harry Saini, and 2 others. 2025. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *ArXiv*, abs/2506.15742.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Midjourney. 2024. Midjourney. https://www.midjourney.com. Text-to-image generative AI model.

Rapidata. 2025. Rapidata image generation alignment dataset. https://huggingface.co/datasets/Rapidata/human-alignment-preferences-images.

Daniel L. Ruderman. 1996. Origins of scaling in natural images. *Vision Research*, 37:3385–3398.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Chameleon Team, Mingda Chen, Jacob Kahn, and Shang-Wen Li. 2024. Chameleon: Mixed-modal early-fusion foundation models. *ArXiv*, abs/2405.09818.

Ping-Rui Tsai, Chi hsiang Wang, Yu-Cheng Liao, and Tzay-Ming Hong. 2025. Three laws of statistical linguistics emerging in images. *ArXiv*, abs/2501.18620.

George Kingsley Zipf. 1932. *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press, Cambridge, MA.