https://doi.org/10.2352/EI.2026.38.12.GENAI-180

A Quantitative Framework for Evaluating Color-Name Understanding in Generative AI Models

Robin Jenkin; NVIDIA; Santa Clara, USA. Vijayalaxmi M, Shailesh Pawale, Francis Fernandes, Ashutosh Naryagol, and Salman Sanadi; KLE Technological University; Hubballi, India

Abstract

With the proliferation of text-to-image generative AI, understanding the fidelity of their output is critical. While these models can generate visually stunning images, their interpretation of nuanced, subjective concepts like color names remains largely unquantified. This paper introduces a systematic framework to evaluate how accurately leading generative AI models (including Flux, Ideogram, Kandinsky, Gemini and Stable Diffusion) understand and reproduce colors from textual prompts. We prompted these models with both one-word (e.g., "blue") and two-word (e.g., "sky blue") color names to generate uniform color fields. The resulting images were analyzed by converting them to the perceptually uniform CIE Lab color space. An adaptive k-means clustering algorithm was employed to extract the dominant color, mitigating issues of non-uniformity in the generated images. By calculating the perceptual color difference using CIEDE2000 (ΔE_{00}) and the chromatic distance (Δab) between the AI-generated colors and standardized ground-truth values, we provide a quantitative benchmark of each model's color accuracy. Our findings reveal that while all models broadly understand the mapping between color names and hue, significant performance variations exist among models, with systematic differences in lightness and chroma reproduction. Per-model analysis reveals a clear hierarchy in chromatic fidelity: Gemini and Flux demonstrate the strongest anchoring, while Kandinsky exhibits striking hue-dependent anisotropy and Stable Diffusion shows the broadest isotropic dispersion. Per-color analysis identifies systematic undersaturation of short-wavelength and high-chroma colors (blue, indigo, magenta) across all models, while warm colors (red, orange, yellow) are generally better grounded. We highlight that results vary significantly across random seeds for the same prompt and model, and that lexical specificity generally—but not universally—improves chromatic grounding. This work provides a robust methodology for auditing and improving color fidelity in future generative models.

Introduction

Text-to-image (T2I) generative artificial intelligence models first emerged in the mid-2010s [22] and have rapidly evolved to produce high-quality images with vivid colors and strong semantic coherence. While aesthetic quality and photorealism are often used to assess these systems, color accuracy provides a more fundamental indicator of semantic grounding. The ability of a model to correctly interpret abstract color terms (e.g., "blue," "sky blue," or "crimson red") reflects the integrity of its language-to-vision mapping.

Human color perception serves as an important reference for evaluating such grounding. Early work by Moroney demonstrated systematic biases in human color naming, including a preference for higher chroma values on digital displays [1]. These findings highlight that even human observers exhibit structured perceptual tendencies. In contrast, generative models inherit statistical biases from large-scale training data. Recent investigations have identified a "narrowing of the color diet," wherein generated images exhibit compressed hue and chroma distributions compared to real-world photography [2, 3]. Additionally, strong object-color priors have been observed, where learned associations between objects and canonical colors (e.g., pumpkins as orange) override explicit chromatic instructions [3].

Recent research has explored color understanding in multimodal and generative systems from several perspectives. Architectural approaches such as ColorPeel aim to disentangle chromatic attributes from shape representations to improve prompt adherence [4]. Other works investigate color compatibility modeling [5], perceptual color-difference learning [6], and statistical palette analysis in visual art [7]. More recently, benchmark-driven evaluations have emerged to assess color reasoning and object-color compositional grounding in complex scenes. These benchmarks typically evaluate color perception within object-rich or semantically entangled contexts.

While such efforts are valuable, they often combine intrinsic chromatic fidelity with object priors, compositional reasoning, and scene semantics. In contrast, the present study isolates the fundamental semantic-to-chromatic mapping by removing object context and photorealistic priors, focusing only on abstract semantic grounding. By restricting prompts to uniform, flat color fields, we directly measure how accurately T2I models ground color-name tokens into perceptually meaningful chromatic outputs.

To this end, we introduce a systematic pipeline that extracts dominant chromatic content from generated images and evaluates perceptual accuracy in CIELab space. We use an adaptive K-Means clustering algorithm to mitigate non-uniform artifacts and compute the CIEDE2000 ΔE_{00} color difference between AI-generated outputs and standardized ground-truth values [8, 17]. This controlled, context-minimized framework enables precise perceptual auditing independent of object-level confounds.

The key contributions of this work are:

- A controlled, context-free evaluation framework for auditing semantic color grounding in T2I generative models.
- An adaptive clustering approach for extraction of dominant chromatic content from non-uniform outputs.
- A comparative perceptual evaluation of five state-of-the-art generative models (Gemini, Flux, Ideogram, Kandinsky, and Stable Diffusion) across one-word and two-word color

prompts.

- A systematic per-model and per-color analysis of chromatic dispersion geometry, revealing distinct failure mechanisms across architectures and hue families.
- A systematic characterization of prompt-level failure modes in semantic-to-chromatic interpretation.

Related Work

Color perception and chromatic grounding in generative models have recently received increasing attention. Early studies investigated statistical properties of color distributions in AI-generated imagery, identifying compressed hue and chroma ranges compared to natural photography [2]. Moroney further explored how color terms are interpreted in Stable Diffusion, highlighting discrepancies between linguistic color labels and chromatic realizations [3].

Architectural approaches such as ColorPeel aim to disentangle chromatic attributes from geometric structure in diffusion models to improve prompt adherence [4]. Other works have examined perceptual color-difference learning [6] and compatibility modeling in large image corpora [5].

More recently, benchmark-driven evaluations have been proposed to assess color perception and reasoning in multimodal systems. Shomer et al. investigate object-color priors and contextual color perception in text-to-image models [13]. ColorBench evaluates color perception, reasoning, and robustness in vision-language models across compositional tasks [14]. GenColorBench introduces a structured benchmark for evaluating color fidelity in text-to-image generation within complex scenes [15]. Additionally, Gomez-Villa et al. analyze how color-name semantics are encoded in vision-language embeddings [16].

While these benchmarks assess compositional reasoning and object-color binding in semantically rich contexts, they inherently couple chromatic fidelity with object priors and scene complexity. In contrast, our work isolates intrinsic semantic-to-chromatic grounding under controlled, context-free conditions. By restricting prompts to uniform color fields and evaluating perceptual ΔE_{00} in CIELab space, we provide a complementary diagnostic framework focused specifically on low-level chromatic accuracy.

Methodology

To evaluate the color fidelity of generative AI models, we developed a quantitative pipeline that extracts dominant chromatic content from generated images and compares it against standardized colorimetric definitions. The framework consists of three stages: prompt generation, image synthesis, and colorimetric analysis.

Prompt Generation and Image Synthesis

A dataset of color prompts was constructed to span both fundamental one-word color terms (e.g., “Blue”, “Red”) and more specific two-word descriptors (e.g., “Sky Blue”, “Midnight Blue”). The prompt formulation was refined through iterative experimentation across models to ensure consistent and interpretable outputs while minimizing semantic ambiguity.

The final standardized prompt used across all models was:

> “A plain, solid colorName background with no textures, patterns, or variations.”

This prompt was selected to suppress object-level semantics and photorealistic priors, thereby isolating intrinsic chromatic grounding. The prompt was evaluated on five generative models: Gemini, Flux, Ideogram, Kandinsky, and Stable Diffusion.

For each color-model pair, a batch of eight images was generated using different random seeds to account for stochastic variability in the diffusion process.

Color Space Transformation

Because the RGB color space is device-dependent and not perceptually uniform, all generated images were converted to the CIE $L^*a^*b^*$ (CIELab) color space prior to analysis. CIELab was selected due to its approximate perceptual uniformity, where Euclidean distances correspond more closely to perceived color differences.

All conversions were performed using the standard D65 illuminant and reference white point to ensure consistency with widely adopted digital imaging standards.

Adaptive Dominant Color Extraction

Generative models frequently fail to produce perfectly uniform color fields, instead introducing gradients, shading, texture artifacts, or unintended imagery. To robustly isolate the intended chromatic content, we implemented an adaptive K-Means clustering algorithm.

Rather than fixing the number of clusters (k) arbitrarily, the algorithm determines k dynamically for each image. The process initializes with $k = 2$ and incrementally increases up to a maximum of $k = 10$. At each step, Euclidean distances between cluster centroids in CIELab space are computed. If the distance between any two centroids falls below a perceptual threshold ($\Delta E_{thresh} = 25$), the algorithm infers that a single perceptual color is being unnecessarily over-segmented. The procedure then terminates, and the centroids from the previous iteration are retained.

After cluster validation, the dominant color is determined using a majority-based criterion:

1. If a single cluster accounts for $\geq 50\%$ of the pixel population, its centroid is selected as the dominant color.
2. If no single cluster reaches this threshold, a weighted mean of the top clusters cumulatively representing 50% of the image area is computed.

This adaptive approach ensures that minor artifacts, boundary noise, or small hallucinated regions do not disproportionately influence the extracted chromatic estimate.

Evaluation Metrics

Color accuracy was quantified using the CIEDE2000 color-difference metric ΔE_{00}, which improves upon the earlier CIE76 formulation by incorporating perceptual weighting functions for lightness, chroma, and hue differences, as well as an interactive term that accounts for perceptual non-uniformities in the blue region [8, 17]. Given the AI-generated dominant color (L_p^*, a_p^*, b_p^*) and the standardized ground-truth color (L_s^*, a_s^*, b_s^*), CIEDE2000 computes a weighted, rotation-corrected distance in CIELab space that more closely reflects human perceptual judgments than the Euclidean CIE76 formulation.

To specifically evaluate chromatic deviation independent of lightness, we also defined the chromatic distance metric Δab:

$$\Delta ab = \sqrt{(a_p^* - a_s^*)^2 + (b_p^* - b_s^*)^2} \quad (1)$$

CIEDE2000 was selected due to its improved perceptual uniformity, direct correspondence to the Just Noticeable Difference (JND) threshold, and widespread adoption in color-science applications [17]. Its corrections for perceptual non-uniformities in the blue region and at low chroma are particularly relevant for evaluating generative model outputs, which frequently exhibit deviations in these areas.

Ground-truth references were established using standardized colorimetric definitions derived from authoritative digital and color-science sources.

Results

Our evaluation reveals a clear dichotomy in generative model behavior: while these systems excel at photorealistic rendering and texture synthesis, they exhibit measurable fragility when tasked with low-level semantic fidelity (i.e., generating a specific, uniform color). The results are presented in two stages: qualitative failure analysis, followed by quantitative evaluation of chromatic accuracy and stability.

Qualitative Analysis: Challenges in Prompt Adherence

Before applying quantitative metrics, we analyzed failure modes observed during dataset generation (Figure 1). These behaviors illustrate structural limitations in semantic-to-chromatic grounding and motivate the controlled evaluation protocol described later.

Semantic Entanglement: Ambiguous color terms such as "Peach" or "Salmon" frequently triggered object generation rather than abstract chromatic fields. Models prioritized learned object associations over the intended color attribute.

Object-Prior Activation: Prompts such as "Forest Green" activated scene-level priors (e.g., foliage imagery), demonstrating strong object-color coupling within the diffusion latent space.

Modality Confusion: Certain models occasionally treated prompts as transcription tasks rather than visual instructions. For example, prompts such as "Blue Square" resulted in rendered text rather than a colored region, indicating interference between text rendering and image synthesis pathways.

Photorealism and Texture Bias: Even when chromatic interpretation was correct, models frequently introduced gradients, lighting effects, or material textures despite explicit instructions requesting uniform fields. These behaviors confirm strong photorealistic priors in T2I systems.

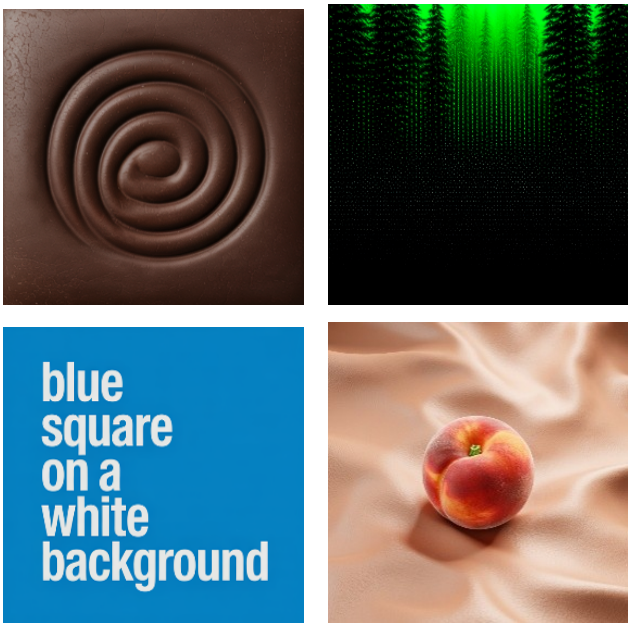


Figure 1. *Representative failure cases observed during prompt evaluation: (a) Texture bias under flat-field instructions, (b) Object-prior activation (scene imagery), (c) Modality confusion with rendered text, and (d) Semantic entanglement where an object is generated instead of the abstract color. These behaviors motivate controlled chromatic evaluation.*

Quantitative Analysis: Colorimetric Accuracy and Stability

Ground-truth $L^*a^*b^*$ values were established using standardized colorimetric definitions from W3C CSS specifications [9], MDN documentation [10], the XKCD color survey [11], and the NIST ISCC–NBS system [12].

Absolute Color Accuracy

Absolute chromatic accuracy was evaluated by computing the mean Δab (chromatic distance) between the extracted dominant color and its corresponding ground-truth reference, isolating hue and saturation deviations from lightness effects.

Figures 2 and 3 visually compare representative color outputs across all five models for one-word and two-word prompts, respectively. Noticeable perceptual deviations are evident in several cases, particularly in short-wavelength and dark chromatic regions.

Figure 4 summarizes mean ΔE_{00} and mean Δab values across all five models, separated by one-word, two-word, and combined prompts.

All models broadly understood the link between color names and hue. However, there are systematic generation differences between models affecting lightness and chroma. Flux and Gemini achieved the lowest overall chromatic deviation, while Kandinsky exhibited the highest overall deviation, driven particularly by elevated error on two-word prompts. Stable Diffusion showed relatively lower chromatic deviation but, as discussed below, exhibited the highest internal variance.

For one-word prompts, Ideogram and Kandinsky showed the greatest chromatic error, whereas Flux and Stable Diffusion demonstrated comparatively lower deviation. Two-word prompts generally reduced chromatic deviation for most models, confirming that lexical specificity constrains chromatic grounding. However, the effect was not universal: Kandinsky showed increased deviation for certain two-word compound color names, suggesting that additional semantic tokens can introduce competing associations rather than stabilizing the chromatic output.

Figure 2. *Comparison of generated dominant color patches across all five models for one-word prompts. The leftmost column represents standardized ground-truth references. Noticeable perceptual deviations are visible across models, particularly for blue, cyan, and magenta.*

Figure 3. *Comparison of generated dominant color patches across all five models for two-word prompts. Two-word descriptors generally yield closer agreement with ground-truth references, though model-specific deviations persist.*

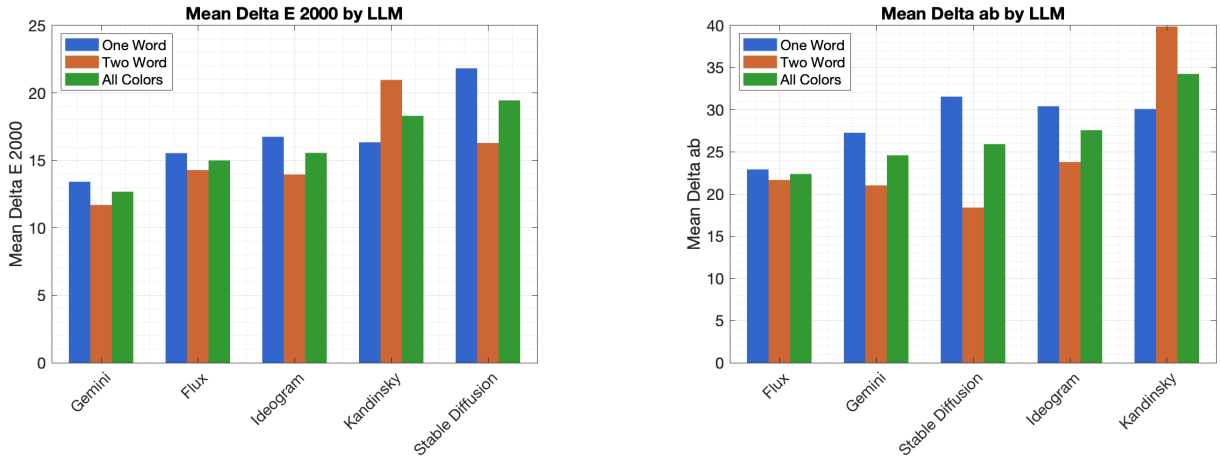


Figure 4. *Left: Mean ΔE_{00} (CIEDE2000 perceptual color difference) per model. Right: Mean Δab (chromatic distance) per model. Both metrics are shown for one-word, two-word, and combined prompts. Lower values indicate closer agreement with standardized ground-truth color definitions. Gemini achieves the lowest error on both metrics, while Kandinsky and Stable Diffusion exhibit the highest deviations.*

Internal Consistency and Spatial Chromatic Dispersion

Beyond absolute accuracy, we evaluated internal stability, defined as the consistency of generated chromatic output across random seeds. For each color–model pair, variance was computed across the batch of eight samples using two complementary metrics: full perceptual variance (ΔE_{00}) using CIEDE2000, and chromatic-only variance in the a^*–b^* plane (Δab). The former captures combined hue and lightness variation, while the latter isolates purely chromatic instability.

Figures 5 and 6 present heatmaps of internal variance for one-word and two-word prompts, respectively.

One-Word Prompts. For basic color terms, Gemini demonstrates comparatively stable behavior across most hues, with predominantly low variance values. Flux also shows generally stable behavior, though with localized spikes in variance for green, purple, blue, and brown.

Ideogram exhibits moderate chromatic spread, particularly in transitional regions. Notably, its Δab variance is elevated for magenta, suggesting partial overlap between adjacent chromatic categories in latent space.

Kandinsky presents heterogeneous stability. While several hues are tightly clustered, blue exhibits exceptionally high variance (ΔE_{00} variance of 353), and magenta also shows elevated dispersion. These patterns indicate that specific chromatic regions are less distinctly separated during diffusion sampling.

Stable Diffusion shows the highest internal variability for one-word prompts. Black (ΔE_{00} variance of 913) and magenta (620) exhibit extreme dispersion, with elevated variance appearing across multiple hue families including both saturated and neutral colors. This suggests greater sensitivity to stochastic sampling and weaker chromatic anchoring.

Two-Word Prompts. The introduction of lexical modifiers alters internal stability in a model-dependent manner (Figure 6). In several cases, added specificity reduces dispersion by constraining the semantic space. Gemini generally exhibits improved stability for compound hues, and Flux maintains relatively coherent outputs across most two-word prompts, with the exception of olive green, which shows elevated variance.

Kandinsky shows an interesting pattern: while most two-word prompts yield stable outputs, sky blue exhibits exceptionally high variance (ΔE_{00} variance of 124), suggesting that certain compound descriptors activate competing chromatic associations.

Stable Diffusion continues to display elevated variance for multiple compound prompts. Lemon yellow exhibits the highest variance across all model–color pairs (ΔE_{00} variance of 455), and deep purple, midnight blue, and goldenrod yellow also show substantial instability, indicating that additional semantic tokens do not uniformly stabilize chromatic output.

To illustrate the nature of high-variance outputs, Figures 7 and 8 present sample generated images for the ten color–model pairs with the highest ΔE_{00} variance for one-word and two-word prompts, respectively. These examples reveal that high variance often corresponds to qualitative failure modes including texture artifacts, object-prior activation, and dramatic chromatic shifts across seeds.

Chromatic vs. Perceptual Variance. Comparison of Δab and ΔE_{00} heatmaps reveals that certain instabilities arise primarily from lightness fluctuations rather than hue drift. This is particularly evident in neutral colors (black, white, and brown), where chromatic dispersion remains limited while perceptual variance increases. Conversely, in blue–purple and high-saturation regions, both metrics increase simultaneously, indicating genuine chromatic instability rather than luminance variation alone.

While heatmaps summarize the magnitude of dispersion, they do not capture its directional structure. To visualize spatial behavior in chromatic space, individual a^*–b^* confidence ellipse diagrams were generated for each color. Each ellipse encloses the dominant chromatic region sampled across seeds, and the star marker indicates the standardized ground-truth reference. Figures 9 and 10 present the complete set of per-color diagrams for one-word and two-word colors, respectively. These visualizations

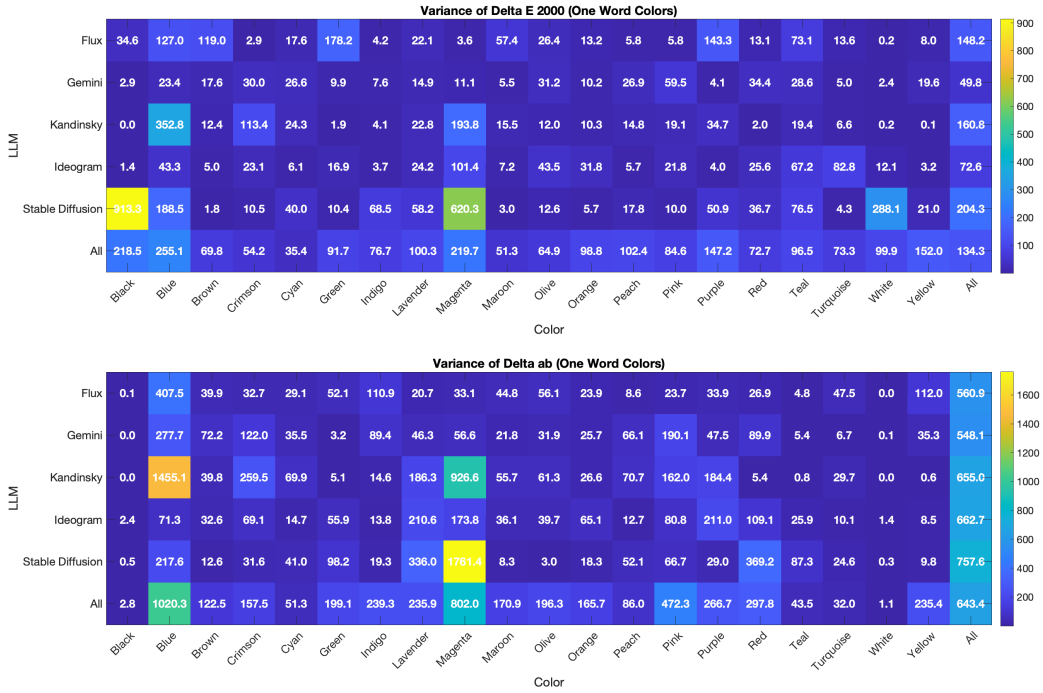


Figure 5. *Heatmaps of internal chromatic variance for one-word prompts. Top: ΔE_{00} (CIEDE2000) variance. Bottom: Chromatic (Δab) variance. Darker regions indicate higher stability; brighter regions indicate increased dispersion across random seeds.*

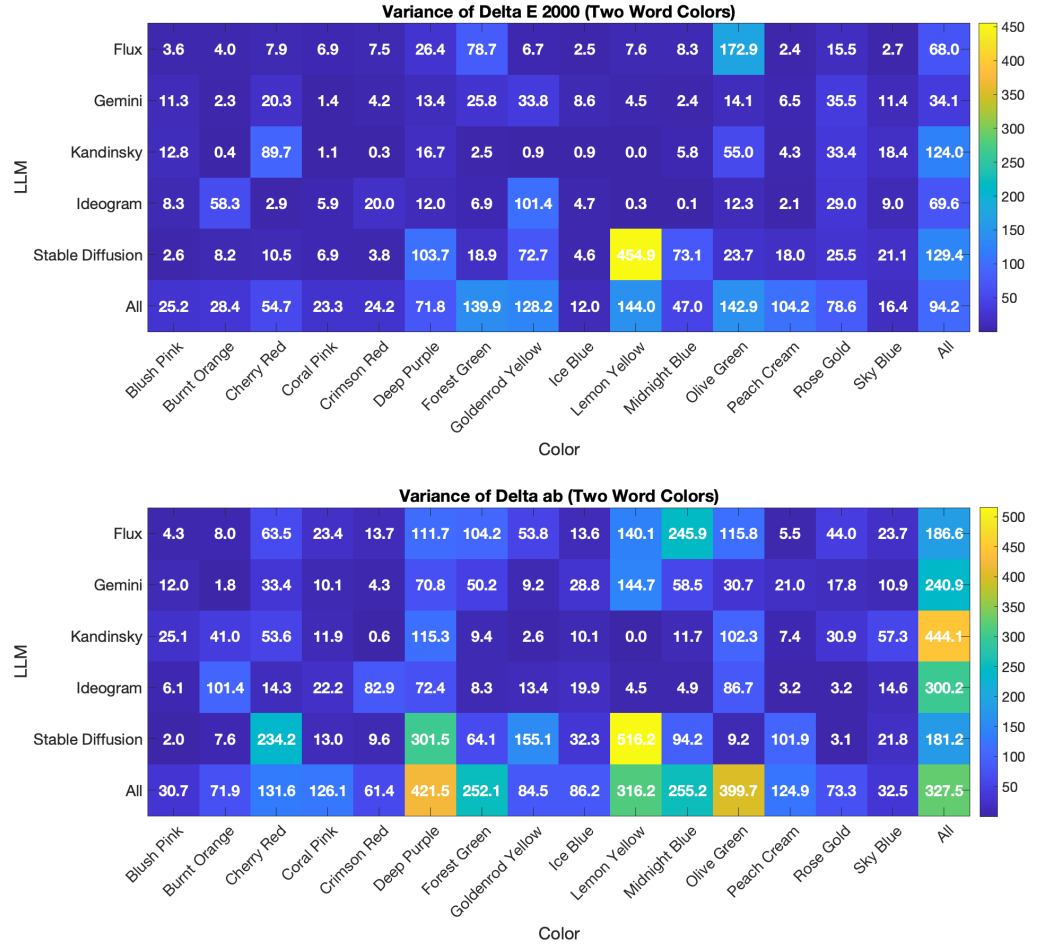


Figure 6. *Heatmaps of internal chromatic variance for two-word prompts. Top: ΔE_{00} (CIEDE2000) variance. Bottom: Chromatic (Δab) variance. Differences relative to one-word prompts reflect the effect of added lexical specificity on chromatic stability.*

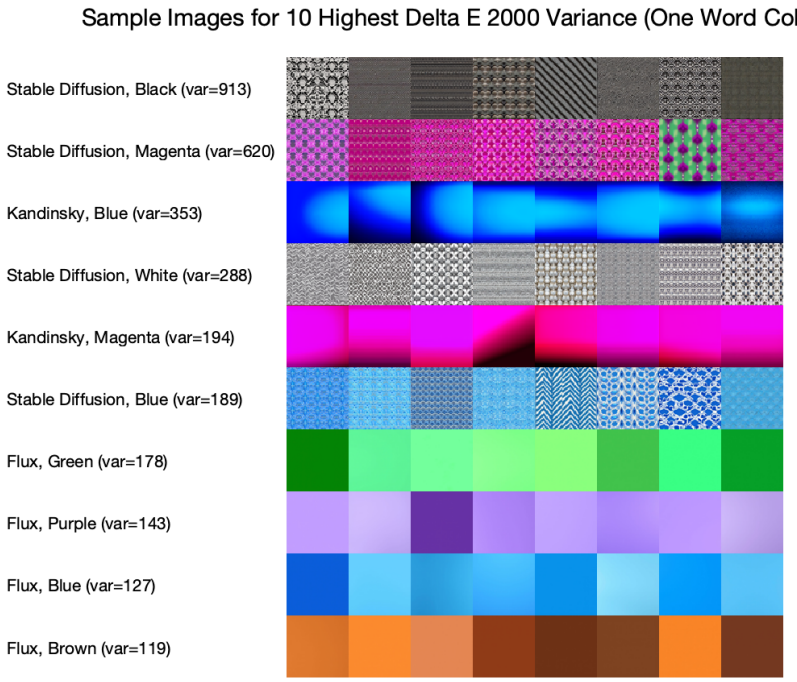


Figure 7. *Sample generated images for the ten color–model pairs with the highest ΔE_{00} variance among one-word prompts. High variance correlates with texture artifacts, non-uniform outputs, and dramatic chromatic shifts across random seeds.*

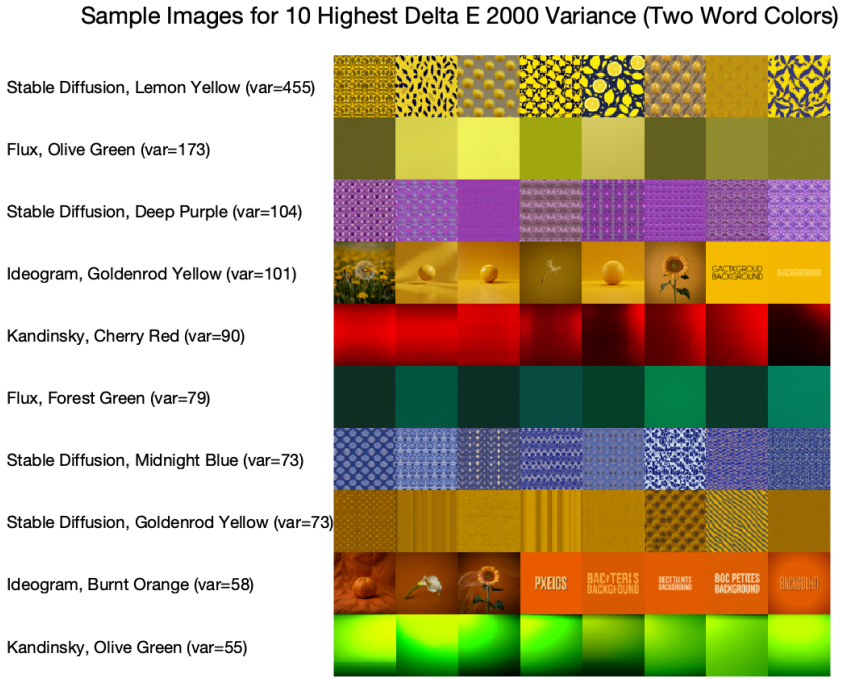


Figure 8. *Sample generated images for the ten color–model pairs with the highest ΔE_{00} variance among two-word prompts. Object-prior activation and modality confusion are visible in several high-variance cases.*

reveal fine-grained dispersion characteristics—including systematic offsets, model-specific directional drift, and anisotropic scatter—that illuminate the geometric structure underlying the variance magnitudes shown in the heatmaps.

Spatial Dispersion Characteristics. The individual a^*–b^* diagrams reveal rich dispersion geometry beyond what the variance heatmaps alone can convey. Gemini consistently produces the most compact ellipses across nearly all colors, centered close to the ground-truth reference, indicating stable, largely isotropic chromatic sampling and strong internal anchoring.

Flux also produces generally compact ellipses, often well-centered near the ground truth, though with localized variance spikes for specific hues. For brown, green, and olive green, Flux exhibits larger ellipses, suggesting weaker chromatic constraint in earth-tone and green spectral regions.

Ideogram shows moderate elongation across several hue families, particularly near chromatic boundaries and at achromatic extremes. Its ellipses for black and white are notably larger than those of Gemini or Flux, indicating reduced stability at low chroma.

Kandinsky exhibits the most distinctive dispersion pattern: strongly anisotropic, highly elongated ellipses for specific hues. The individual plots reveal that this elongation is directional rather than random—Kandinsky's ellipses for crimson, cherry red, purple, pink, and olive stretch along narrow axes in the a^*–b^* plane, sometimes extending far from the ground truth. This pattern suggests that the diffusion process traverses extended trajectories through neighboring chromatic regions during sampling, producing structured drift rather than isotropic scatter. Notably, for other colors such as yellow and red, Kandinsky produces very tight, well-centered clusters, indicating strongly hue-dependent chromatic anchoring.

Stable Diffusion produces the broadest and most isotropic (circular) ellipses across the widest range of hues. In contrast to Kandinsky's elongated directional drift, Stable Diffusion's dispersions tend toward symmetric spread, suggesting general stochastic instability rather than structured chromatic drift. The large, offset ellipses observed for lemon yellow, peach cream, turquoise, and magenta indicate a combination of internal variance and systematic chromatic bias.

Together, the heatmaps and spatial ellipses reveal that internal chromatic behavior differs not only in magnitude but also in geometric structure. Some models exhibit compact, centered clusters (Gemini, Flux), while others demonstrate directional drift (Kandinsky) or broader isotropic dispersion (Stable Diffusion), reflecting fundamental differences in how chromatic information is represented and sampled within latent space.

Systematic Analysis by Model

The individual per-color a^*–b^* diagrams enable a more granular characterization of each model's chromatic behavior, revealing systematic patterns that transcend individual color categories.

Gemini. Gemini demonstrates the most consistent chromatic grounding across the evaluated color set. Its ellipses are almost uniformly compact and well-centered near the ground-truth reference for both one-word and two-word prompts. The primary limitation is a tendency toward slight desaturation: for green, the generated output is noticeably less chromatically vivid than the standardized reference, and for magenta the ellipse center falls at lower a^* values than the ground truth. These observations suggest that Gemini's latent color space may compress extreme chroma values while preserving hue fidelity.

Flux. Flux generally achieves strong chromatic accuracy with moderate internal variance. Its ellipses are typically centered near the ground truth, though they are somewhat larger than Gemini's for several hue families. Flux shows localized instability in earth tones and greens: brown, olive green, and forest green produce notably larger ellipses. For warm colors such as red and orange, Flux maintains good centering with moderate spread. The overall profile suggests reliable hue grounding with weaker chroma constraint in certain spectral regions.

Ideogram. Ideogram exhibits moderate performance, with a distinctive weakness at achromatic extremes. For black, its ellipse extends substantially into negative b^* territory, and for white, elevated variance is visible despite the nominally simple chromatic target. For saturated hues, Ideogram produces moderately sized ellipses with slight directional elongation near chromatic boundaries in the purple–magenta region. Its two-word prompt performance is generally more stable, suggesting that added semantic context helps constrain its latent representation.

Kandinsky. Kandinsky presents the most heterogeneous behavior. Its defining characteristic is extreme anisotropy: for specific hue families, it produces dramatically elongated ellipses extending along narrow axes in the a^*–b^* plane. The most extreme cases include crimson (extending to $a^* > 120$, $b^* > 110$), cherry red, purple, and pink, where ellipses span a substantial fraction of the visible chromaticity range. This behavior is consistent with weakly bounded diffusion trajectories in specific chromatic regions [18]. However, Kandinsky's performance is not uniformly poor: for yellow and red it produces some of the tightest clusters of any model, centered very close to ground truth. This dichotomy suggests strongly hue-dependent chromatic anchoring, with well-separated latent representations for some colors but overlapping or poorly constrained representations for others.

Stable Diffusion. Stable Diffusion exhibits the broadest overall dispersion. Its ellipses are consistently large across multiple hue families, and the dispersion geometry tends toward isotropic rather than directional, distinguishing it from Kandinsky's anisotropic drift. Particularly high variance is observed for black (ΔE_{00} variance of 913), magenta, lemon yellow (ΔE_{00} variance of 455), and turquoise. Beyond variance, Stable Diffusion also shows systematic offset—its ellipse centers are frequently displaced from the ground-truth reference, indicating a combination of internal instability and persistent chromatic bias. This dual characteristic makes Stable Diffusion the least reliable model for precision color generation among those evaluated.

Systematic Analysis by Color

Complementary to the per-model analysis, examining behavior across models for individual colors reveals color-dependent

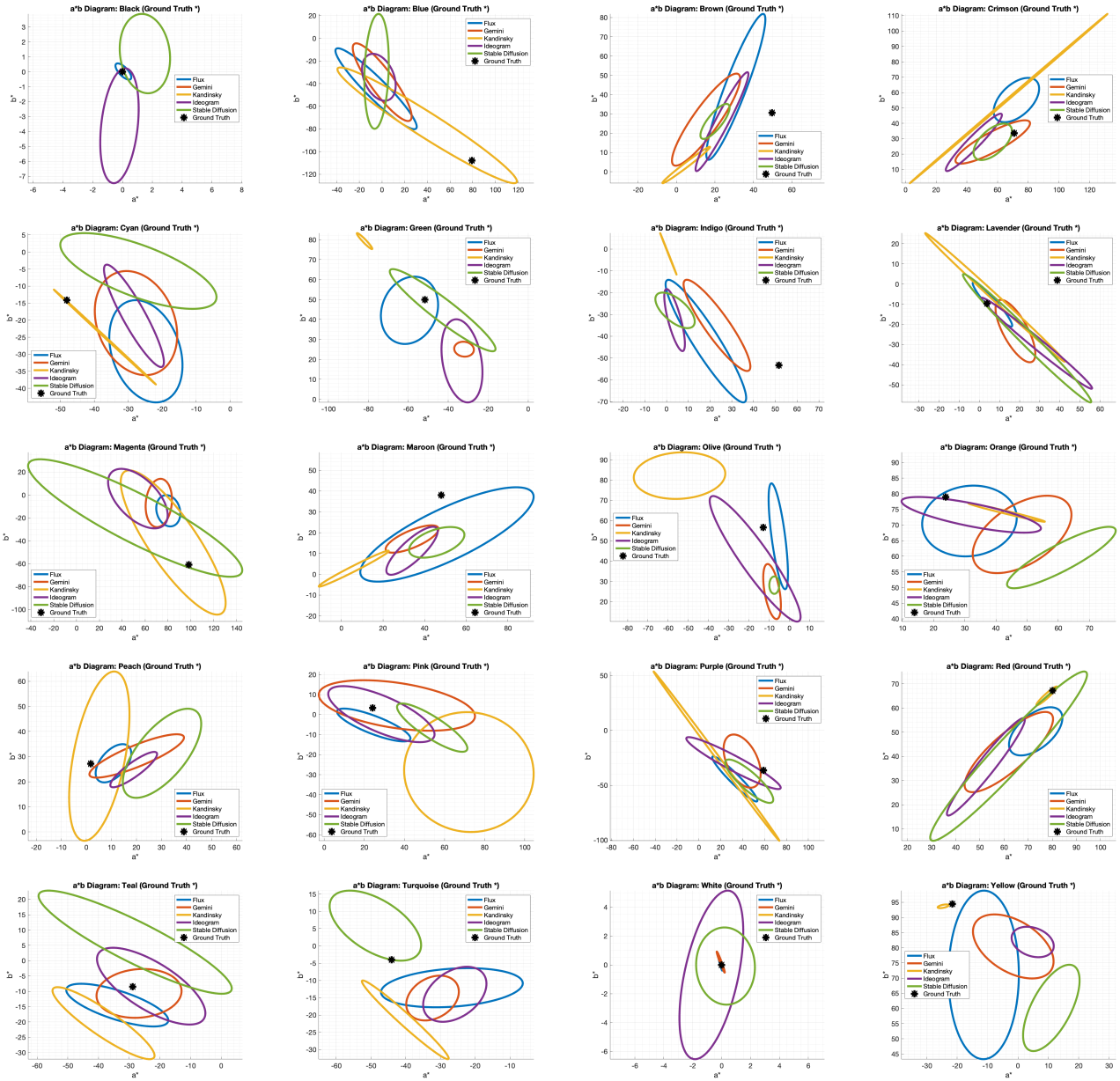


Figure 9. *Individual a^*–b^* confidence ellipse diagrams for all one-word colors. Each subplot shows the spatial dispersion of dominant chromatic samples across random seeds for all five models, with the star marker indicating the standardized ground-truth reference. Ellipse size reflects variance magnitude while orientation reveals directional chromatic drift. Notable patterns include universal undersaturation for blue and indigo, extreme anisotropy from Kandinsky for crimson and purple, and broad isotropic dispersion from Stable Diffusion across multiple hues.*

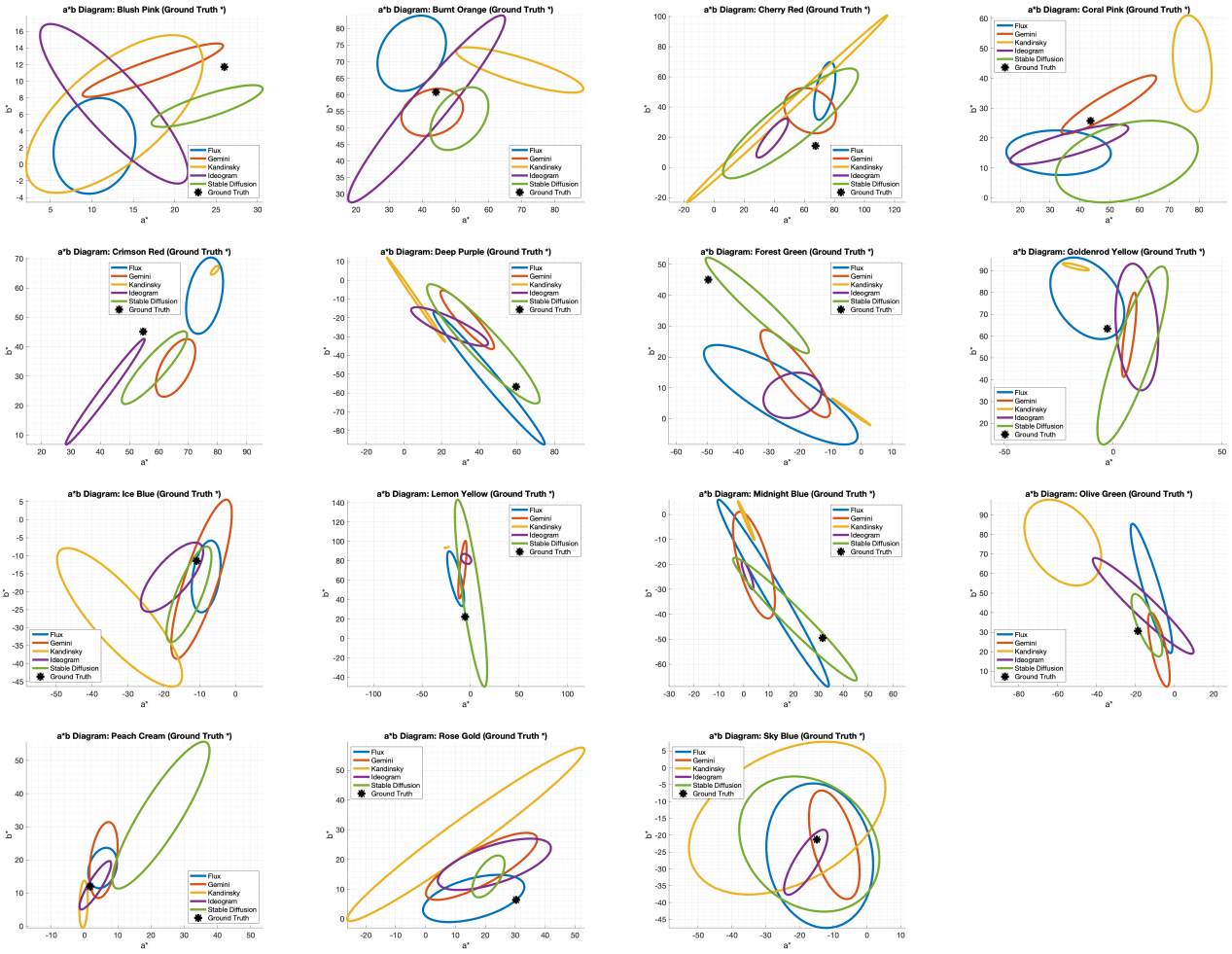


Figure 10. *Individual a^*–b^* confidence ellipse diagrams for all two-word colors. Compound descriptors yield a range of behaviors from tight clustering (e.g., burnt orange, coral pink) to dramatic model-dependent divergence (e.g., crimson red with Kandinsky, sky blue with elevated variance). The additional semantic specificity of two-word prompts generally constrains dispersion but can also activate competing chromatic associations in certain model–color pairs.*

patterns in semantic-to-chromatic grounding.

Achromatic Colors (Black, White). Black and white should produce tight clustering near the origin of the a^*–b^* plane. Gemini, Flux, and Kandinsky achieve this for both colors. However, Ideogram and Stable Diffusion show substantially elevated variance, with ellipses extending several units from the origin (Figure 9). This suggests that these models do not strongly anchor the concept of "achromatic" in latent space, allowing stochastic variation to introduce chromatic contamination.

Blue, Indigo, and Deep Purple. These short-wavelength colors expose a systematic failure across all models. For blue, the standardized ground truth lies at approximately ($a^* = 75, b^* = -107$)—a highly saturated region of the chromaticity plane. However, all five models produce ellipses clustered near ($a^* \approx -10$ to $20, b^* \approx -30$ to -60), dramatically undershooting both components (Figure 9). This universal offset indicates that the training data representation of "blue" maps to a desaturated, less chromatically extreme region than the colorimetric standard. Indigo and deep purple show analogous patterns, with all models clustering at substantially lower a^* values than the reference. These findings are consistent with the "narrowing of the color diet" described by Wang et al. [2], where generative models compress the chromaticity range relative to ground-truth standards. The effect is particularly acute for highly saturated blue-region colors, suggesting that the training corpora underrepresent extreme chromaticity values associated with pure spectral colors.

Magenta and Pink. Magenta (ground truth at approximately $a^* = 95, b^* = -60$) represents one of the most chromatically extreme test targets. All models fail to reach this level of saturation, producing ellipses centered at substantially lower a^* values. Stable Diffusion and Kandinsky show the largest variance, while Gemini and Flux produce tighter but still offset clusters. The consistent undersaturation across all models suggests a fundamental limitation in the gamut of the learned chromatic representation, possibly reflecting the prevalence of sRGB-gamut training data [2].

Green, Cyan, and Teal. Green (ground truth at approximately $a^* = -55, b^* = 50$) shows substantial model-dependent behavior. Flux and Stable Diffusion produce large ellipses in the correct quadrant, while Gemini generates a tight but notably desaturated cluster. Kandinsky's ellipse is displaced toward very high b^* values, indicating a shift toward yellow-green. For cyan and teal, all models shift toward less negative a^* than the reference, indicating systematic undersaturation in the green direction.

Red, Crimson, and Crimson Red. Red is one of the best-handled colors across all models. The ground truth at approximately ($a^* = 78, b^* = 67$) falls within a region that all models approximate reasonably well, with Kandinsky producing particularly tight clustering. However, the addition of modifiers introduces instability: crimson triggers an extremely elongated ellipse for Kandinsky, extending along a diagonal axis to extreme chromaticity values. This illustrates how lexical modifiers can destabilize rather than refine chromatic grounding for some model–color pairs [20].

Yellow and Lemon Yellow. Yellow (ground truth at approximately $a^* = -22, b^* = 94$) is well-handled by Kandinsky, which produces a tight cluster near the reference. Flux generates a large ellipse, and Stable Diffusion is displaced toward positive a^* values (indicating a reddish shift). For lemon yellow, Stable Diffusion exhibits the highest variance of any model–color pair (ΔE_{00} variance of 455), with its ellipse extending from $b^* \approx -40$ to over $b^* = 140$ (Figure 10).

Brown, Maroon, and Olive. Earth tones are generally reproduced in the correct quadrant but with notable offsets. Flux shows the highest variance for brown. Olive is particularly challenging: Kandinsky's ellipse is displaced dramatically upward to $b^* \approx 80$–90, while other models cluster at more negative a^* values with moderate dispersion.

Compound Descriptors and Modifier Effects. The two-word color analysis reveals that lexical modifiers do not uniformly improve grounding. For forest green, midnight blue, and deep purple, all models produce ellipses offset from ground truth by comparable magnitudes to their one-word counterparts (Figure 10). Sky blue shows elevated variance for Kandinsky, and cherry red produces extreme elongation for the same model. These patterns indicate that added semantic tokens can activate competing associations in some architectures, introducing rather than reducing chromatic ambiguity [20].

Discussion

The results show that color generation in text-to-image models is not uniformly reliable, even under tightly controlled flat-field prompts. While all evaluated systems are capable of producing visually plausible outputs, their chromatic grounding varies across hue categories, prompt types, and sampling conditions. The per-model and per-color analyses presented above reveal several integrated observations that extend beyond individual color–model pairs.

First, all models broadly understood the link between color names and hue. The generated outputs were generally in the correct hue family, confirming that the fundamental semantic-to-chromatic mapping is intact across architectures. However, there are systematic generation differences between models affecting lightness and chroma. As shown by the per-model chromatic distance analysis (Figure 4), models exhibit distinct chromatic "signatures" that persist across color categories.

Second, the per-model analysis reveals a clear hierarchy in chromatic fidelity and stability. Gemini demonstrates the most consistent performance, with compact, well-centered ellipses across nearly all colors. Flux follows closely, with slightly larger variance but comparable centering. Ideogram and Kandinsky occupy an intermediate tier with pronounced hue-dependent performance—Kandinsky in particular exhibits a striking dichotomy between tightly anchored colors (yellow, red) and dramatically unstable ones (crimson, purple, pink). Stable Diffusion shows the weakest overall performance, with both the largest variance and the most systematic offset from ground truth, consis-

tent with prior observations of compressed color distributions in diffusion-based systems [2, 3].

Third, the improvement observed with two-word prompts was model-dependent. Adding lexical modifiers generally reduced absolute chromatic deviation and, in several cases, improved internal stability. This suggests that increased semantic specificity constrains the latent representation and narrows the range of plausible chromatic interpretations [20]. However, the effect was not universal. For some models and color families—most notably Kandinsky with sky blue and certain red-derived compounds—additional descriptors introduced competing semantic associations rather than stabilizing the output.

Fourth, the per-color analysis reveals that chromatic failures are not randomly distributed but cluster in specific regions of the a^*–b^* plane. Short-wavelength colors (blue, indigo, deep purple) and high-chroma colors (magenta, cyan) exhibit the largest universal offsets from ground truth, with all models undershooting the target chromaticity. This systematic undersaturation across architectures points to a shared limitation—likely rooted in the statistics of training data, which are dominated by sRGB-gamut imagery and may underrepresent the extreme chromaticity values of pure spectral colors [2]. By contrast, warm colors (red, orange, yellow) are generally better grounded, suggesting that training corpora more densely sample these regions of color space.

Fifth, the distinction between chromatic variance (Δab) and full perceptual variance (ΔE_{00}) reveals that instability is not always driven by hue drift. In neutral tones, variation often arises from lightness fluctuations rather than shifts in chromatic coordinates. In contrast, short-wavelength and high-saturation regions show genuine dispersion in the a^*–b^* plane, indicating weaker separation between adjacent chromatic categories in latent space.

Sixth, the geometry of dispersion provides mechanistic insight. Kandinsky's strongly anisotropic, elongated ellipses—particularly for crimson, cherry red, and purple—suggest that its diffusion trajectories traverse extended paths through neighboring chromatic regions during the reverse process, consistent with weakly bounded sampling in specific directions of latent space [18]. In contrast, Stable Diffusion's broader, more isotropic dispersions suggest a more general sensitivity to stochastic noise, analogous to the "mode exploration" behavior described in diffusion model literature [19]. These distinct dispersion geometries imply different underlying failure mechanisms despite similar aggregate variance magnitudes.

Seventh, the results vary significantly for the same prompt and model across different random seeds. The high-variance sample images (Figures 7 and 8) demonstrate that individual generations can deviate dramatically from the batch centroid, with some seeds producing texture artifacts, object imagery, or entirely different hues despite identical prompts.

Importantly, these observations were obtained under minimal compositional complexity. The prompts were explicitly designed to isolate color generation without object semantics or scene structure. The persistence of chromatic variance under such constraints indicates that the observed limitations are intrinsic to semantic-to-chromatic mapping rather than artifacts of compositional interference.

Overall, the findings suggest that while modern generative models have strong perceptual rendering capabilities, their internal representation of color categories remains uneven. Lexical specificity improves grounding in many cases, but chromatic stability varies across hue families and models in a structured manner. These differences likely reflect underlying training distributions, latent compression strategies, and the manner in which color tokens are embedded within multimodal representations.

Relation to Existing Benchmarks

Recent benchmarks such as ColorBench [14] and GenColorBench [15] evaluate color reasoning, robustness, and object-color compositional grounding within semantically rich scenes. These frameworks provide valuable insight into multimodal reasoning and attribute binding; however, they inherently couple chromatic fidelity with object priors and scene complexity.

In contrast, the present study isolates low-level semantic-to-chromatic mapping under context-minimized conditions. By restricting prompts to uniform color fields and evaluating perceptual ΔE_{00} in CIELab space, we directly measure intrinsic color-name grounding independent of compositional confounds. This controlled evaluation complements object-centric benchmarks by providing a diagnostic baseline for chromatic accuracy.

Implications for Imaging Applications

The observed deviations have practical implications for color-critical workflows, including brand identity design, digital illustration, accessibility design, and scientific visualization. In such applications, perceptual color accuracy is not merely aesthetic but functionally significant. The elevated ΔE_{00} values observed across models indicate that current off-the-shelf T2I systems are not yet reliable for tasks requiring strict chromatic precision without post-processing or model fine-tuning.

Taken together, compositional benchmarks and controlled perceptual audits provide complementary perspectives on color understanding in generative models. While reasoning-focused benchmarks evaluate high-level attribute binding, controlled chromatic evaluation reveals intrinsic perceptual alignment limitations. Both dimensions are necessary for a comprehensive assessment of color competence in generative AI systems.

Conclusion

This paper introduced a controlled, CIELab-centric framework for auditing semantic color fidelity in text-to-image generative models. By moving beyond subjective visual inspection to quantitative evaluation using CIEDE2000 ΔE_{00}, chromatic distance Δab, and adaptive dominant-color extraction, we systematically analyzed how five state-of-the-art models interpret abstract color terms. Individual per-color a^*–b^* dispersion diagrams enabled granular characterization of both model-specific and color-specific patterns.

Our results demonstrate that all evaluated models broadly understand the mapping between color names and hue, but intrinsic chromatic grounding remains inconsistent. The per-model analysis reveals a clear hierarchy: Gemini demonstrates the strongest chromatic anchoring with consistently compact, well-centered ellipses; Flux follows closely with slightly greater variance; Ideogram and Kandinsky exhibit pronounced hue-dependent performance with localized instabilities; and Stable Diffusion shows the broadest dispersion and most systematic offset. Kandinsky's behavior is particularly distinctive, with extreme anisotropic drift for specific hue families (crimson, purple, pink)

contrasting with tight clustering for others (yellow, red), suggesting strongly hue-dependent latent space structure.

The per-color analysis reveals that chromatic failures are not randomly distributed. Short-wavelength colors (blue, indigo, deep purple) and high-chroma colors (magenta, cyan) show the largest universal offsets from ground truth across all models, consistent with the "narrowing of the color diet" in training data. Warm colors (red, orange, yellow) are generally better grounded, likely reflecting denser sampling of these regions in training corpora. One-word prompts generally exhibit higher perceptual deviation than two-word descriptors, though this effect is model-dependent, and certain compound names can introduce competing semantic associations.

These findings have direct implications for color-critical applications, including brand identity systems, accessibility design, digital illustration, and scientific visualization. The elevated chromatic deviations observed across models indicate that current off-the-shelf T2I systems are not yet reliable for precision color workflows without post-processing or fine-tuning.

Importantly, the comparatively strong performance observed in Gemini and Flux suggests that accurate chromatic grounding is achievable within existing architectures. Future research should focus on disentanglement strategies [4], improved color-aware text encoders, training objectives that explicitly respect perceptual color-space relationships, incorporation of wider-gamut training data to address the systematic undersaturation of short-wavelength colors, and the potential for explicitly teaching models about color through specialized fine-tuning [21].

By complementing object-centric color benchmarks with controlled perceptual auditing and detailed per-color chromatic analysis, this work contributes a foundational diagnostic tool for evaluating and improving color understanding in generative AI systems.

References

[1] N. Moroney, "Unconstrained web-based color naming experiment," *Proc. SPIE 5008, Color Imaging VIII: Processing, Hardcopy, and Applications*, (2003).

[2] Y. Wang, M. A. Webster, and D. S. Joyce, "Color statistics of images created by generative AI," *J. Opt. Soc. Am. A*, 42, B76-B80 (2025).

[3] N. Moroney, "Color Terms and Stable Diffusion," *Proc. IS&T Color and Imaging Conference (CIC32)*, pp. 84–88 (2024).

[4] M. A. Butt, K. Wang, J. Vazquez-Corral, and J. van de Weijer, "ColorPeel: Color Prompt Learning with Diffusion Models via Color and Shape Disentanglement," *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, vol. 15065, pp. 456–472, Springer (2024).

[5] P. O'Donovan, A. Agarwala, and A. Hertzmann, "Color Compatibility From Large Datasets," *ACM Transactions on Graphics (Proc. SIGGRAPH)*, (2011).

[6] H. Chen, Z. Wang, Y. Yang, Q. Sun, and K. Ma, "Learning a Deep Color Difference Metric for Photographic Images," *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2023).

[7] J. Romero, L. Gomez-Robledo, and J. L. Nieves, "Computational color analysis of paintings for different artists of the XVI and XVII centuries," *Color Research & Application*, 43(3), 296–303 (2018).

[8] G. Sharma (Ed.), *Digital Color Imaging Handbook*, CRC Press, Boca Raton, FL, 2003.

[9] C. Lilley et al., "CSS Color Module Level 3," *World Wide Web Consortium (W3C) Recommendation*, (2022).

[10] Mozilla Developer Network, "CSS Named Colors," *MDN Web Docs*, (2025).

[11] R. Munroe, "The XKCD Color Survey," *XKCD*, (2010).

[12] K. L. Kelly and D. B. Judd, "Color: Universal Language and Dictionary of Names," *National Bureau of Standards (NIST) Special Publication 440*, (1976).

[13] S. Shomer-Chai, W. Peng, B. Hariharan, and H. Averbuch-Elor, "Color Bind: Exploring Color Perception in Text-to-Image Models," *arXiv preprint arXiv:2508.19791*, (2025).

[14] Y. Liang et al., "ColorBench: Can VLMs See and Understand the Colorful World? A Comprehensive Benchmark for Color Perception, Reasoning, and Robustness," *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, (2025).

[15] M. A. Butt et al., "GenColorBench: A Color Evaluation Benchmark for Text-to-Image Generation Models," *arXiv preprint arXiv:2510.20586*, (2025).

[16] A. Gomez-Villa et al., "Color Names in Vision-Language Models," *arXiv preprint arXiv:2509.22524*, (2025).

[17] M. R. Luo, G. Cui, and B. Rigg, "The Development of the CIE 2000 Colour-Difference Formula: CIEDE2000," *Color Research & Application*, 26(5), 340–350, (2001).

[18] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 6840–6851, (2020).

[19] Y. Song et al., "Score-Based Generative Modeling through Stochastic Differential Equations," *Proc. International Conference on Learning Representations (ICLR)*, (2021).

[20] R. Rassin, E. Hirsch, D. Glickman, S. Ravfogel, Y. Goldberg, and G. Chechik, "Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment," *Advances in Neural Information Processing Systems (NeurIPS)*, (2023).

[21] D. Podell et al., "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis," *Proc. International Conference on Learning Representations (ICLR)*, (2024).

[22] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative Adversarial Text to Image Synthesis," *Proc. International Conference on Machine Learning (ICML)*, 48, 1060–1069, (2016).

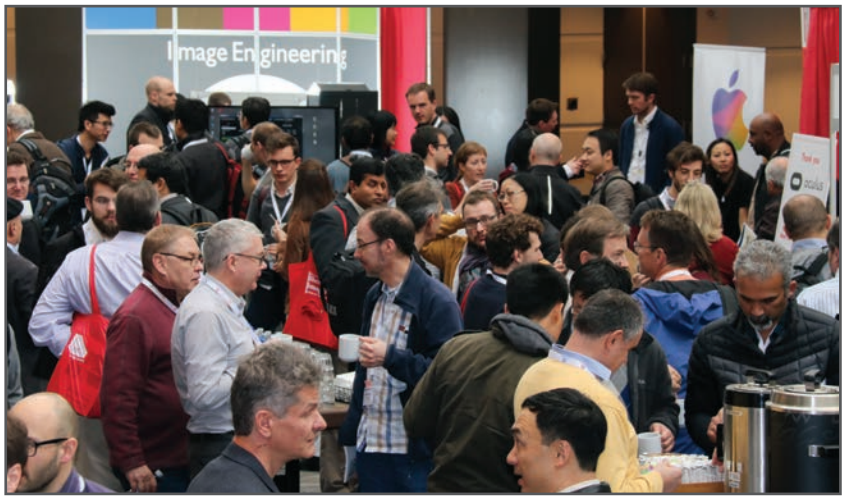
Author Biography

Robin Jenkin received a BSc(Hons) in Photographic and Electronic Imaging Science (1995) and his PhD (2001) in the field of image science from the University of Westminster. He also holds a M.Res Computer Vision and Image Processing from University College London (1996). Robin is a Fellow of The Royal Photographic Society, UK, and Executive Vice President of Society for Imaging Science and Technology. Robin is secretary of the IEEE P2020 Image Quality for Autonomous Vehicles working group. At NVIDIA Robin is a distinguished engineer and models image quality for autonomous vehicle and other applications. He is a Visiting Professor at University of Westminster within the Computer Vision and Imaging Technology Research Group and co-author of the 10th ed. "The Manual of Photography", Focal Press.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org