ColorConceptBench: A Benchmark for Probabilistic Color-Concept Understanding in Text-to-Image Models

Anonymous ACL submission

Abstract

While text-to-image (T2I) models have advanced considerably, their capability to associate colors with implicit concepts remains underexplored. To address the gap, we introduce *ColorConceptBench*, a new human-annotated benchmark to systematically evaluate color-concept associations through the lens of probabilistic color distributions. *ColorConceptBench* moves beyond explicit color names or codes by probing how models translate 1,281 implicit color concepts using a foundation of 6,369 human annotations. Our evaluation of seven leading T2I models reveals that current models lack sensitivity to abstract semantics, and crucially, this limitation appears resistant to standard interventions (e.g., scaling and guidance). This demonstrates that achieving human-like color semantics requires more than larger models, but demands a fundamental shift in how models learn and represent implicit meaning.

1 Introduction

Recent advances in text-to-image (T2I) generation can synthesize high-fidelity images that align semantically with text prompts, demonstrating sophisticated control over object composition (Huang et al., 2023; Ghosh et al., 2023) and spatial relationships (Bakr et al., 2023; Gokhale et al., 2022). Yet a significant challenge remains: these models often struggle to accurately understand and render color semantics, particularly in tasks requiring precise color rendering and complex attribute binding (Liang et al., 2025; Butt et al., 2025). Enabling generative models to capture the nuanced color-concept associations is a critical yet underexplored challenge in achieving true semantic alignment with human perception.

To evaluate how well T2I models understand color-concept associations, existing approaches typically adopt a two-stage *generate-and-evaluate* pipeline, as illustrated in Figure 1. These methods (Lin et al., 2024; Tsai et al., 2025; Butt et al., 2025) rely on explicit color specifications, such as color names (e.g., '*green*') or color codes (e.g., '*#00FF00*') within the text prompt (e.g., '*A clipart of {color} forest*') during the generation phase. The subsequent evaluation relies on deterministic verification by measuring the distance between the rendered colors and a single, pre-defined ground truth, treating color alignment as a pass/fail check against a reference value.

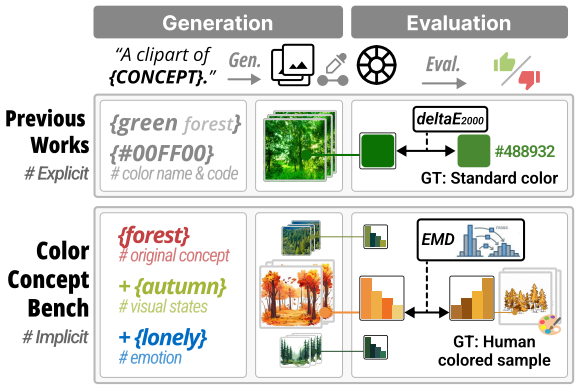

Figure 1: Unlike explicit color matching (top), *ColorConceptBench* evaluates implicit semantic alignment using probabilistic color distributions (bottom).

However, color is more than a physical attribute, but also a carrier of semantic information in human cognition. In creative practice, users rarely specify precise color codes; instead, they rely on descriptions of visual states (e.g., '*autumn*') or emotional atmospheres (e.g., '*lonely*') to guide generation in prompts (Hou et al., 2025). Existing benchmarks lack this semantic depth, failing to assess the advanced semantic alignment required for interpreting such implicit visual concepts. Moreover, humans develop color associations through perceptual experience, which manifests as a *probabilistic distribution of color expectations* rather than a fixed, one-to-one mapping (Schloss, 2024). In contrast, existing approaches are fundamentally flawed, as they reduce a nuanced semantic landscape to a single point, discarding essential information about

both diversity and associative intensity.

To bridge these gaps, we introduce *ColorConceptBench*, a new benchmark for evaluating the probabilistic alignment of implicit color semantics in T2I models. We construct an human-grounded dataset comprising 1,281 concepts and 6,369 human-annotated samples, collected from a controlled sketch colorization task performed by 168 designers to ensure collective consensus. Accordingly, *ColorConceptBench* provides a comprehensive evaluation framework that utilizes distribution-based metrics (e.g., EMD) to quantify the alignment between model-generated color profiles and human ground truth. We conduct extensive evaluations of seven leading T2I models. Our analysis reveals a shortcoming of model insensitivity to abstract concepts: while models can reproduce object colors, they consistently struggle to infer appropriate colors from abstract variants (e.g., visual states and emotions). Furthermore, we show that simply scaling up models or increasing guidance strength does not resolve this gap. The findings underscore implicit color alignment as a persistent challenge.

Our contributions are summarized as follows:

- **Human-Grounded Color-Concept Association Benchmark:** We introduce a new benchmark grounded in professional designer annotations, including 6,369 sketch colorizations for 1,281 color concepts, to quantify the probabilistic gap between AI-generated color profiles and human color-concept associations.
- **Probabilistic Evaluation Protocol:** We establish an evaluation protocol that includes both probabilistic and deterministic feature alignment, providing a granular framework to guide future research in improving semantic color controllability.
- **Systematic Evaluation and Insights:** We conduct a comprehensive evaluation of leading T2I models across varied concepts, styles, and guidance scales. Our analysis reveals that current models lack sensitivity to implicit semantics, a limitation that remains resistant to both parameter scaling and stronger guidance.

2 Related Work

Color-Concept Association. Color is a fundamental semantic channel, conveying emotional tones and cultural associations beyond mere visual appearance (Wierzbicka, 1990). The accurate association of color with concept is critical for applications such as graphic design (Jahanian et al., 2017) and interior design (Hou et al., 2024). Research supporting these applications is grounded in empirical data collection, where prior work has pursued through two primary approaches. The first involves human annotation or preference ranking, where participants directly identify or order colors for given concepts (Rathore et al., 2019; Mukherjee et al., 2024; Volkova et al., 2012). However, this method is often constrained by small scale and labor intensity. The second approach overcomes this limitation by employing automated or generative methods to extract color–concept associations at scale (Bahng et al., 2018; Hou et al., 2025).

While generative model-based approaches show promise, they are largely constrained to colors inherent to object appearance, such as '*red*' for '*apple*' or '*green*' for '*grass*' (Setlur and Stone, 2015). This focus limits their semantic diversity, failing to capture the implicit color associations essential for abstract concepts. For instance, when prompted with implicit concepts like '*lonely*' or '*festive*', models often produce inconsistent or semantically misaligned colors, failing to reflect the emotional or cultural palettes that humans intuitively expect; see Appendix Figure 7 for example. To address this limitation, our work introduces a benchmark centered on implicit color concepts, including abstract visual states and emotional associations, enabling a more comprehensive evaluation of high-level semantic alignment in T2I generation.

Color Evaluation in Generative Models. Color is a critical dimension for ensuring both visual realism and semantic alignment in T2I models. Existing color evaluations primarily focus on attribute binding (Butt et al., 2024; Samin et al., 2025), assessing whether a model correctly associates a specified color with a target object, using metrics like CLIPScore (Radford et al., 2021) or VQA accuracy (Liang et al., 2025). A more granular line of work performs fine-grained color verification, testing pixel-level rendering of specific color names or hexadecimal codes (Bahng et al., 2018; Butt et al., 2025; Tsai et al., 2025). Recent studies have started exploring implicit concepts like emotion and culture in generated images, such as to evaluate the cultural competence (Senthilkumar et al., 2024) and emotional control (Dang et al., 2025).

These approaches typically reduce to a deter-

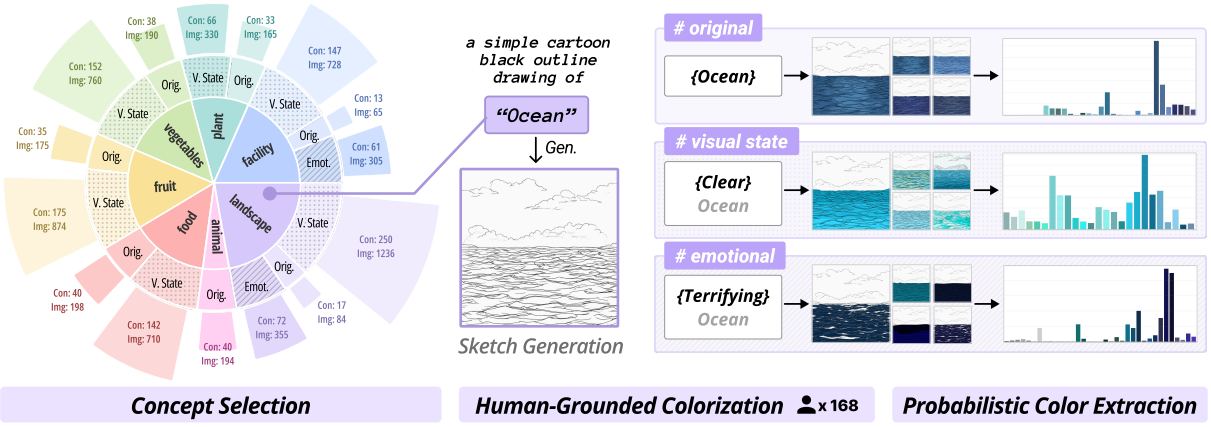

Figure 2: Dataset Statistics and Construction Pipeline. An overview of the hierarchical concept distribution and our three-stage construction process: concept selection, human-grounded colorization, and probabilistic color extraction.

ministic check, using metrics like Euclidean distance in RGB space, to determine if a generated object's dominant color matches a singular reference value. This deterministic approach, however, is misaligned with the nature of color semantics that are probabilistic distributions, not deterministic point values (Schloss, 2024). To address this limitation, our approach shifts from singular reference colors to human-grounded color distributions. We crowdsource representative colorization from multiple designers for each target concept and extract their collective color profiles, capturing the varied ways humans visualize abstract ideas. This allows us to evaluate alignment using distribution-based metrics, such as Earth Mover's Distance (EMD) (Rubner et al., 2000), to measure how closely a model's generated color distribution matches the human perceptual ground truth.

3 *ColorConceptBench*

ColorConceptBench is a benchmark for evaluating probabilistic color-concept understanding in T2I models. Formally, let $\mathcal{C}$ denote a set of semantic concepts (e.g., '*lonely*', '*festive*'). For each concept $c \in \mathcal{C}$, we define a *human-grounded color* as a probability distribution over a color space Ω:

$$P_{\mathrm{H}}(x \mid c), \quad x \in \Omega,$$

where Ω is a suitable color space. This distribution is constructed empirically from a curated dataset of human visual annotations of c. The construction of *ColorConceptBench* is illustrated in Figure 2.

3.1 Concept Selection

Color requires a visual carrier to be realized in a generated image. As such, we begin by selecting a set of target concepts that correspond to concrete objects, which serve as these carriers. We source these object concepts from the THINGS dataset (Hebart et al., 2019), a large and systematically curated collection of visually grounded entities. To ensure linguistic relevance and common usage, we apply a subsequent filter based on word frequency from the COCA[1] to ensure relevance and coverage.

To evaluate models' understanding of implicit color concepts in practical contexts, we further define a set of descriptive adjectives that modify entity-based concepts. Based on the characteristics of each entity category, we first source relevant adjectives from lexical databases and combine them with the original concepts. Detail links are provided in Table 4. The selected attribute dimensions are divided into two categories:

- ***Visual State*** describes attributes perceived through observation, such as '*polluted*' or '*clear*' water, and '*unripe*' plants.
- ***Emotional*** covers less-explored affective dimensions, including moods like '*cozy*' or '*lonely*'. We restrict *emotional* assignments to *facility* and *landscape* entities, as these contexts naturally support affective interpretation.

Then, all the adjectives are collected and refined through iterative review with collaborating designers to ensure both linguistic and visual validity. After combining the original objects with these adjectives, we obtained a final set of 1,281 unique distributed across 8 distinct categories. The statistics of concepts in each object type and attribute dimensions are listed in Figure 2, with the list of concepts provided in Appendix B.1.

[1] https://www.english-corpora.org/coca/

3.2 Human-Grounded Colorization

Sketch Generation. After defining the concept set, we generate simple sketches (rather than real images) for each concept, to minimize stylistic variability, reduce background noise, and isolate color as the primary variable of interest. Using Qwen-Image (Wu et al., 2025a) and Stable Diffusion 3.5 Medium (Esser et al., 2024), we generate five sketches per concept with each model. A collaborating artist then manually selects the cleanest and most unambiguous sketch for each concept to ensure clarity in subsequent color annotation.

Human Colorization. Next, we invite professional designers to colorize the selected sketches. Each concept is independently colored by five designers, who are instructed to base their color choices on real-world, intuitive associations drawn from everyday experience with the concept, rather than on personal artistic style. We recruit designers for their acute color sensitivity, which reduces annotation noise and yields more reliable human ground truth. The full instructions and designer demographics are documented in Appendix B.3. Finally, we review all submissions and remove clear outliers to ensure that the collected color data reflects stable, semantically grounded human associations. The quality control protocol and validation results are detailed in Appendix B.4. In total, we recruit 168 designers for colorization, each coloring 40 to 50 sketches. Following our quality control protocol, we retained 6,369 high-quality colored results for the final dataset (see Appendix Figure 6 for example).

3.3 Probabilistic Color Extraction

Segmentation. We first use Grounding DINO (Liu et al., 2024) to identify the bounding box of the target concept. The detected bounding box is then passed to the Segment Anything Model (SAM) (Kirillov et al., 2023), which generates a binary mask of the target concept. This step ensures that the subsequent analysis is derived strictly from pixels belonging to the concept.

Quantization and Refinement. To capture the primary visual impression, we extract representative colors in the CIELAB space. However, since a single concept often exhibits multiple color distributions (e.g., '*apple*' can be distinctively red or green), simply averaging all samples would result in inaccurate representations. To address this, we implement an adaptive grouping strategy:

1. *Perceptual Grouping.* We first cluster images into distinct visual groups based on the CIE ΔE2000 distance between their dominant colors, ensuring that visually disparate modes are processed separately.

2. *Refined Merging.* Within each group, we perform a merging process where color centers indistinguishable to the human eye are combined. The final distribution $P_H(x \mid c)$ over the color space Ω is constructed by aggregating these refined color distribution, weighted by the population of each visual group.

Implementation details are provided in Appendix C.3.

4 Experiment

4.1 Models

We evaluate the performance of seven well-known open-source T2I models on the benchmark. These models include Stable Diffusion (SD) XL (Podell et al., 2023), SD 3 (Esser et al., 2024), SD 3.5 (Esser et al., 2024)from the stability AI, Flux.1-dev (Labs, 2024), Qwen-Image (Wu et al., 2025a), OmniGen2 (Wu et al., 2025b), and SANA-1.5 (Xie et al., 2024).

4.2 Implementation

For each concept $c \in \mathcal{C}$, we construct a diverse image set by varying inference parameters to ensure robustness. To determine if color associations are style-dependent, we generate images across two styles (*natural* and *clipart*) using standardized templates. Furthermore, to analyze the impact of generation constraints, we sample images across 7 distinct Classifier-Free Guidance (CFG) scales. For each combination concept, style, and guidance scale, we generate 5 independent samples at 1024×1024 resolution. The prompt templates and detailed generation configurations are provided in Appendix C.1 and C.2. Next, to derive the probability distribution $P_{\mathrm{M}}(x \mid c)$ from these generated images, we employ the identical probabilistic color extraction pipeline described in Sect. 3.3

4.3 Metric

4.3.1 Probabilistic Distribution Alignment

We employ the following metrics to quantify the statistical and perceptual alignment between these

distributions. For computational convenience, let $p \in \mathbb{R}^K$ and $q \in \mathbb{R}^K$ denote the discrete probability vectors corresponding to P_{H} and P_{M} respectively, over the K bins of the UW71 color space (Hu et al., 2022; Rathore et al., 2019).

Pearson Correlation Coefficient (PCC). PCC measures the linear correlation between the color probability distribution perceived by human and those learned by the model.

$$PCC(p,q) = \sum_{k=1}^{K} \frac{(p_k - \bar{p})(q_k - \bar{q})}{\sqrt{(p_k - \bar{p})^2}\sqrt{(q_k - \bar{q})^2}}, \quad (1)$$

where $\bar{p}$ and $\bar{q}$ denote the mean probabilities of the respective distributions.

Earth Mover's Distance (EMD). EMD accounts for the perceptual distance between color distributions. It computes the minimum cost required to transform p into q, using a ground distance matrix D, where $d_{k_1k_2}$ represents the CIELAB Euclidean distance between color bins k_1 and k_2:

$$EMD(p,q) = \min_{f} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} f_{k_1k_2} d_{k_1k_2}, \quad (2)$$

subject to flow constraints $\sum_{k_2} f_{k_1k_2} = p_{k_1}$, $\sum_{k_1} f_{k_1k_2} = q_{k_2}$, and $f_{k_1k_2} \geq 0$.

Entropy Difference (ED). To evaluate whether a generative model captures the complexity and diversity of human color concept associations, we compute the absolute difference in Shannon entropy between the two distributions, as:

$$ED(p,q) = |\sum_{k=1}^{K} p_k \log(p_k) - \sum_{k=1}^{K} q_k \log(q_k)|. \quad (3)$$

4.3.2 Deterministic Feature Alignment

Similar to previous works relying on deterministic metrics, we further extract dominant color from the underlying distributions to evaluate features alignments.

Dominant Color Accuracy (DCA). We identify the "dominant color" as the bin with the highest probability mass in the aggregated distribution. For a specific concept c, let $k_H^{(c)} = \arg\max_k(p_k)$ and $k_M^{(c)} = \arg\max_k(q_k)$ be the indices of the peak color bins for humans and the model. We calculate the accuracy over the set of concepts N:

$$DCA = \frac{1}{|\mathcal{C}|} \sum_{c \in |\mathcal{C}|} \mathbb{1}(k_H^{(c)} = k_M^{(c)}), \quad (4)$$

where $k_H^{(c)}$ and $k_M^{(c)}$ denote the dominant colors, and $\mathbb{1}(\cdot)$ is the indicator function. This metric strictly assesses the model's ability to precisely capture the representative color of the concept.

Hue Angular Difference (ΔHue). To explicitly evaluate chromatic alignment, we compute the angular difference between the mean hue angles of the dominant color. For a concept c, let $\theta_H^{(c)}$ and $\theta_M^{(c)}$ be the hue angles associated with the dominant bins $k_H^{(c)}$ and $k_M^{(c)}$. The metric is defined as the shortest angular distance between these two angles, averaged over all concepts:

$$\Delta Hue = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \min(|\Delta\theta^{(c)}|, 360^\circ - |\Delta\theta^{(c)}|), \quad (5)$$

where $\Delta\theta^{(c)} = \theta_H^{(c)} - \theta_M^{(c)}$. This metric quantifies the divergence in the overall hue direction.

4.4 Results

Probabilistic Distribution Alignment. Table 1 presents the probabilistic distribution alignment of T2I models across concept categories and visual styles. All models show moderate alignment with human color perception, as indicated by PCC exceeding 0.60 for most concepts. Notably, for a given model, alignment is higher for Original concept than for its variants Visual state and Emotional, and higher for Clipart than for Natural imagery. Among all models, Sana-1.5 (4.8B) achieves state-of-the-art performance, especially on Clipart images, where it surpasses the runner-up, Flux.1-dev, especially based on the EMD metric. This indicates that Sana-1.5, despite its smaller parameter count compared to larger models like Qwen-Image, excels at mapping textual semantics to corresponding color distributions across all concepts.

Deterministic Feature Alignment. Table 2 presents the deterministic feature alignment of T2I models across concept categories and visual styles. Similarly, for a given model, alignment is consistently highest for the Original concept, followed by its Visual state and Emotional variants, and higher for Clipart than for Natural imagery. Most models achieve a ΔHue below 30°, particularly for Clipart style in the Original concept, indicating their output closely matches the human-designated dominant hue. For Natural images, SD 3 and 3.5 demonstrate superior color fidelity, achieving the lowest ΔHue errors overall, which suggests a more precise grasp of the target hue than competing models.

Model	Original						Visual state						Emotional					
	Natural			Clipart			Natural			Clipart			Natural			Clipart		
	PCC↑	EMD↓	ED↓	PCC↑	EMD↓	ED↓	PCC↑	EMD↓	ED↓	PCC↑	EMD↓	ED↓	PCC↑	EMD↓	ED↓	PCC↑	EMD↓	ED↓
Stable Diffusion 3 (2B)	**0.696**	29.971	0.596	0.717	29.301	0.593	0.683	29.913	0.530	0.675	31.775	0.608	0.671	**30.674**	0.534	0.659	31.629	0.548
Stable Diffusion 3.5 (2.5B)	0.661	**29.717**	0.590	0.677	29.516	0.625	0.642	30.783	0.536	0.655	30.691	0.579	0.642	31.647	0.537	0.652	32.083	0.513
Stable Diffusion XL (2.6B)	0.665	31.143	0.599	0.698	28.796	0.677	0.653	**28.529**	**0.492**	0.721	**27.320**	0.612	0.588	30.782	**0.479**	0.688	29.592	0.665
Sana-1.5 (4.8B)	0.678	31.797	**0.540**	**0.743**	26.548	**0.477**	**0.696**	29.733	0.554	**0.722**	28.137	**0.484**	**0.689**	31.398	0.642	0.715	**26.731**	**0.495**
OmniGen2 (4.8B)	0.629	32.111	0.595	0.681	31.367	0.587	0.611	34.055	0.591	0.663	33.794	0.599	0.545	35.958	0.550	0.565	34.489	0.569
Flux.1-dev (12B)	0.682	31.146	0.585	0.733	**26.396**	0.512	0.673	30.138	0.563	0.710	28.642	0.516	0.667	32.650	0.673	**0.719**	29.154	0.500
Qwen-Image (20B)	0.661	30.646	0.637	0.704	29.776	0.643	0.636	31.976	0.660	0.673	30.749	0.628	0.652	30.714	0.572	0.675	32.147	0.574

Table 1: Comparison of *probabilistic distribution alignment* of different T2I models across concepts and styles. The best and second best results in each column are marked in **bold** and underlined, respectively.

Model	Original				Visual state				Emotional			
	Natural		Clipart		Natural		Clipart		Natural		Clipart	
	DCA↑	Δ Hue↓	DCA↑	Δ Hue↓	DCA↑	Δ Hue↓	DCA↑	Δ Hue↓	DCA↑	Δ Hue↓	DCA↑	Δ Hue↓
Stable Diffusion 3 (2B)	0.169	**32.702**	0.169	28.965	0.118	39.773	0.113	38.135	0.120	**45.303**	0.158	**42.529**
Stable Diffusion 3.5 (2.5B)	0.145	32.997	0.159	30.464	0.116	**38.910**	0.128	37.211	0.120	53.300	0.120	46.725
Stable Diffusion XL (2.6B)	0.070	41.810	0.178	33.974	0.113	47.954	0.135	38.903	0.090	61.622	0.120	56.198
Sana-1.5 (4.8B)	0.131	33.900	0.182	26.691	0.149	40.419	**0.193**	35.996	**0.173**	53.752	**0.165**	48.839
OmniGen2 (4.8B)	0.164	37.107	0.164	30.514	0.114	43.223	0.135	40.697	0.135	56.238	0.098	49.612
Flux.1-dev (12B)	**0.178**	35.566	**0.192**	**26.626**	**0.154**	39.879	0.169	36.941	0.128	55.648	0.120	58.065
Qwen-Image (20B)	0.150	35.260	0.154	30.447	0.125	42.077	0.121	**34.279**	0.098	53.863	0.105	50.857

Table 2: Comparison of *deterministic feature alignment* of different T2I models across concepts and styles. The best and second best results in each column are marked in **bold** and underlined, respectively.

Qualitative Results. Figure 3 presents qualitative examples generated by different models across various concepts and styles. These examples align with the quantitative findings discussed earlier. For instance, outputs from Qwen-Image, which shows the weakest human alignment in the metrics, demonstrate a stronger reliance on form and composition than on color matching, particularly for the Visual State and Emotional concepts.

4.5 Human Judgment

To validate the reliability of our evaluation metrics, we conducted a perceptual study to measure their consistency with human judgment. We randomly sampled 150 concepts and generated outputs using three representative models: Sana, SDXL, and OmniGen. This experiment employs a pairwise comparison protocol hosted on a Gradio interface (Abid et al., 2019). To ensure statistical robustness, we invite 36 participants, where each pair was evaluated by six independent annotators. For each trial, participants are presented with the human-annotated ground truth alongside the images generated by two randomly paired models. Participants are asked to determine which model's output aligned more closely with the human ground truth.

We quantify alignment between our metrics and human preference using Pearson (r) and Spearman (ρ) correlation coefficients, and agreement ratio.

Type	Metric	Pearson	Spearman	Agreement
Probabilistic	PCC	0.4155	0.4020	55.33%
	EMD	**0.5334**	**0.5468**	**62.44%**
	ED	0.2571	0.2589	49.78%
Deterministic	DCA	0.3075	0.3040	13.56%
	ΔHue	0.2252	0.2595	39.11%

Table 3: Validation of metric alignment with human judgment. Metrics based on probabilistic distribution, especially EMD, align more closely with human judgment than those based on deterministic features.

Table 3 presents the results, which reveal that the adopted distribution-based metric, especially EMD, exhibits a significantly higher correlation with human judgment compared to deterministic metrics commonly employed in prior studies, which typically assess performance based on single-color accuracy. Detailed information about human judgment can be found in Appendix D.

5 Discussion

5.1 Abstract Concepts are Difficult to Color

A trend observed across most models is the performance degradation from concrete to abstract concepts, namely the *Visual State* and *Emotional* semantic modifiers. Models generally achieve their best scores in the *Original* category (simple nouns). However, performance consistently drops when processing *Visual State* modifiers, and reaches its nadir in the *Emotional* category. This suggests that

Figure 3: Qualitative comparison of color-concept association across different text-to-image models. Colors shift for base nouns (e.g., 'cabin') and modified concepts involving visual states (e.g., 'cozy') or emotions (e.g., 'lonely'), across both natural and clipart styles, shown with color distribution and dominant colors with sample number.

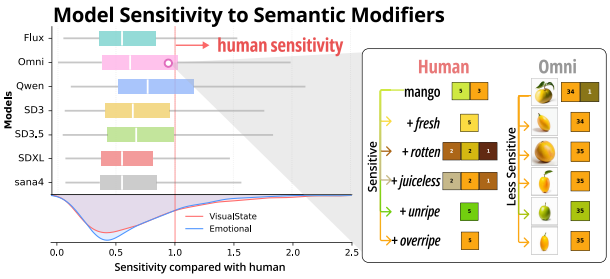

Figure 4: Models consistently exhibit lower color shift magnitudes than the human baseline (left), conservatively prioritizing intrinsic object colors over modifier-induced adjustments (right).

while models effectively retrieve fixed object-color associations (e.g., 'apple' → red), they struggle to process more implicit concepts.

To further investigate this gap, we evaluate the model's semantic sensitivity by measuring the color shift induced by modifiers. Specifically, we calculate the EMD between the color distribution of a base concept (e.g., 'Mango') and its modified version (e.g., 'Rotten Mango'). A larger EMD signifies a larger distributional difference, indicating a stronger response to the modifier. We compare its shift intensity against the human baseline. For instance, human designers exhibit varying degrees of intensity depending on the modifier (e.g., a drastic palette shift for 'rotten' vs. a subtle adjustment for 'fresh'). Ideally, models should exhibit a shift intensity comparable to humans; a significantly lower intensity indicates that the model fails to update the color distribution in response to the semantic cue. However, we find that models are consistently **under-sensitive**, as shown in Figure 4 (left). Quantitatively, for *Visual State* modifiers (see Appendix Table 8), the average color shift produced by models is only ∼**78%** of the human shift intensity. This sensitivity drops further to ∼**69%** for *Emotional*

modifiers. This indicates a conservative tendency: models prioritize the intrinsic color of the noun (e.g., keeping a mango yellow) rather than making the generative adjustments required by the adjective.

By cross-referencing sensitivity magnitude with visual outputs, we identify three distinct response behaviors: *(1) Semantic Inertia (Low Sensitivity).* Models like OmniGen fail to override strong object priors, exhibiting negligible response to modifiers. For instance, distinct prompts like '*cozy*' and '*lonely*' cabin produce nearly identical color distributions (Figure 3), indicating a failure to initiate the necessary distributional shift. *(2) Semantic Over-Correction (High Sensitivity / Drift).* Finally, models such as Qwen exhibit excessive sensitivity that can lead to semantic drift (See Appendix Figure 14). While their shift magnitude rivals that of humans, they often lack control. For instance, when generating a "polluted lake," the model may completely overwrite the water's blue tones with mud colors, whereas human annotators typically retain the blue hue while introducing grayish tones. Here, high sensitivity reflects a failure to preserve the subject's identity, rather than accurate adaptation and balancing. *(3) Precise Adaptation (Balanced Sensitivity).* In ideal cases, models like SDXL and Sana achieve small but semantically accurate shifts. For the '*rotten apple*' case (Figure 3), SDXL effectively shifts the palette from canonical red to brownish-green decay without losing the object's identity. Similarly, Sana demonstrates a subtle yet directionally aligned color shift. This demonstrates a balance: these models interpret the modifier as a specific physical property update rather than a generic style transfer.

5.2 Inefficacy of Scaling and Guidance

Inefficacy of Stronger Guidance. While Classifier-Free Guidance (CFG) is typically increased to enhance prompt fidelity, our results indicate it generally fails to improve semantic color alignment. As shown in Figure 5, models like OmniGen, SD 3, and SD 3.5 exhibit a counter-intuitive degradation in performance, characterized by a general upward trend in EMD scores and a simultaneous decline in PCC. In contrast, SDXL, Sana, and Qwen demonstrate remarkable stability, maintaining flat trends regardless of guidance strength. This suggests that implicit color binding behaves as a fixed intrinsic capability that resists improvement through inference-time tuning, unlike spatial adherence, which is highly responsive to guidance.

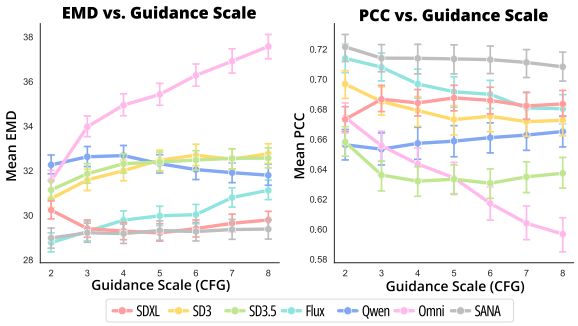

Figure 5: Impact of Guidance Scale. Increasing the CFG scale generally leads to higher EMD and lower PCC across models (worse alignment).

Absence of Scaling Law. We observe no correlation between model parameter and semantic color alignment, even among models with similar backbones (e.g., DiT variants like Sana, and Qwen). Notably, SDXL (2.6B) outperforms larger models, including OmniGen (4.8B) and Qwen (20B). This indicates that scaling model capacity is insufficient for implicit semantic binding. Instead, factors such as training data distribution and architectural inductive biases (e.g., U-Net vs. Autoregression) likely play a more decisive role. Future work could investigate data augmentation or synthetic data generation techniques that strengthen the link between abstract color concept association.

6 Conclusion

In this work, we introduce *ColorConceptBench* , a benchmark to evaluate color-concept association, an important yet under-explored aspect of text-to-image generation. After evaluating seven T2I models on 1,281 concepts with over 6,369 human annotations, our study reveals that current T2I models still struggle to associate concepts with human-expected colors. Specifically, we find that models fail to maintain color-concept associations ability as semantic complexity increases. Moving forward, we plan to extend our investigation to cross-cultural contexts, exploring how human color-concept associations vary across different backgrounds and whether T2I models can capture these culturally specific semantic nuances. We envision this benchmark as a foundation to foster future advancements in semantic-aware color generation.

Limitations

There are several limitations of this work. 1) *Demographic and Cultural Scope.* Our human annotation data is predominantly collected from a single cultural region. Since color symbolism and emotional associations are culturally dependent, our current findings represent a specific geo-cultural distribution. Consequently, while this provides a baseline, the observed color-concept associations may not generalize to global contexts with distinct cultural color associations. 2) *Scope of Concept and Semantic Modifiers.* While our study extensively investigates abstract semantic modifiers, we restrict the base concepts to tangible objects with intrinsic visual properties. We cover seven classes of natural and artificial objects, explicitly excluding purely abstract subjects (e.g., '*happy*' or '*freedom*'). This ensures that we measure color shifts relative to a stable structural anchor, thereby avoiding the unbounded visual variability inherent in the generation of abstract concepts. Regarding the semantic modifier, another limitation is the omission of culturally specific cues. Color-concept associations are often culture-dependent rather than universal. For instance, the concept "wedding" is strongly associated with Red in many Eastern cultures (symbolizing luck and joy), whereas it is predominantly linked to White in Western traditions (symbolizing purity).

Ethics Statement

We acknowledge the broader ethical implications of generative AI. The study was approved by the authors' institution's ethics board, with all participants providing informed consent in data annotation process and human judgment. Regarding the human evaluation, we ensured that all data was collected anonymously with informed consent obtained from participants. We confirm that our dataset contains no personally identifiable information (PII) and has been screened to be free of offensive content or societal biases. Consequently, the release of this benchmark poses no foreseeable harm to society and does not facilitate the generation of malicious content like deepfakes. All models used are publicly available, and we will release our code and dataset to ensure reproducibility after peer review period.

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Hyojin Bahng, Seungjoo Yoo, Wonwoong Cho, David Keetae Park, Ziming Wu, Xiaojuan Ma, and Jaegul Choo. 2018. Coloring with words: Guiding image colorization through text-based palette generation. In *Proceedings of the European Conference on Computer Vision*, pages 431–447.

Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. 2023. HRS-Bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20041–20053.

Muhammad Atif Butt, Alexandra Gomez-Villa, Tao Wu, Javier Vazquez-Corral, Joost Van De Weijer, and Kai Wang. 2025. Gencolorbench: A color evaluation benchmark for text-to-image generation models. *arXiv preprint arXiv:2510.20586*.

Muhammad Atif Butt, Kai Wang, Javier Vazquez-Corral, and Joost van de Weijer. 2024. Colorpeel: Color prompt learning with diffusion models via color and shape disentanglement. In *European Conference on Computer Vision*, pages 456–472. Springer.

Shengqi Dang, Yi He, Long Ling, Ziqing Qian, Nanxuan Zhao, and Nan Cao. 2025. Emoticrafter: Text-to-emotional-image generation based on valence-arousal model. *arXiv preprint arXiv:2501.05710*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning*.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152.

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. 2022. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*.

Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. 2019. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, 14(10):e0223792.

Yihan Hou, Manling Yang, Hao Cui, Lei Wang, Jie Xu, and Wei Zeng. 2024. C2Ideas: Supporting creative

interior color design ideation with a large language model. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, page 172:1 – 172:18.

Yihan Hou, Xingchen Zeng, Yusong Wang, Manling Yang, Xiaojiao Chen, and Wei Zeng. 2025. GenColor: Generative color-concept association in visual design. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 544:1–544:19.

Ruizhen Hu, Ziqi Ye, Bin Chen, Oliver van Kaick, and Hui Huang. 2022. Self-supervised color-concept association via image colorization. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):247–256.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747.

Ali Jahanian, Shaiyan Keshvari, SVN Vishwanathan, and Jan P Allebach. 2017. Colors–messengers of concepts: Visual design mining for learning color semantics. *ACM Transactions on Computer-Human Interaction*, 24(1):1–39.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Black Forest Labs. 2024. Flux. https://github.com/black-forest-labs/flux.

Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. 2025. Colorbench: Can vlms see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. *arXiv preprint arXiv:2504.10514*.

Jieru Lin, Danqing Huang, Tiejun Zhao, Dechen Zhan, and Chin-Yew Lin. 2024. Designprobe: A graphic design benchmark for multimodal large language models. *arXiv preprint arXiv:2404.14801*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, pages 38–55. Springer.

Kushin Mukherjee, Timothy T Rogers, and Karen B Schloss. 2024. Large language models estimate fine-grained human color-concept associations. *arXiv preprint arXiv:2406.17781*.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PmLR.

Ragini Rathore, Zachary Leggon, Laurent Lessard, and Karen B Schloss. 2019. Estimating color-concept associations from image statistics. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1226–1235.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121.

Ahnaf Mozib Samin, M. Firoz Ahmed, and Md Mushtaq Shahriyar Rafee. 2025. ColorFoil: Investigating color blindness in large vision and language models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 294–300.

Karen B Schloss. 2024. Color semantics in human cognition. *Current Directions in Psychological Science*, 33(1):58–67.

Nithish Kannen Senthilkumar, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: Cultural competence in text-to-image models. *Advances in Neural Information Processing Systems*, 37:13716–13747.

Vidya Setlur and Maureen C Stone. 2015. A linguistic approach to categorical color assignment for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):698–707.

Danielle Albers Szafir. 2017. Modeling color difference for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):392–401.

Sung-Lin Tsai, Bo-Lun Huang, Yu Ting Shen, Cheng Yu Yeo, Chiang Tseng, Bo-Kai Ruan, Wen-Sheng Lien, and Hong-Han Shuai. 2025. Color me correctly: Bridging perceptual color spaces and text embeddings for improved diffusion generation. *arXiv preprint arXiv:2509.10058*.

Svitlana Volkova, William B Dolan, and Theresa Wilson. 2012. Clex: a lexicon for exploring color, concept and emotion associations in language. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 306–314.

Anna Wierzbicka. 1990. The meaning of color terms: semantics, culture, and cognition. *Cognitive Linguistics*, pages 99–150.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, and 1 others. 2025a. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, and 1 others. 2025b. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and 1 others. 2024. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*.

A Appendix: Statements

Reproducibility Statement. To facilitate reproducibility, we will make the entire dataset, source code, and scripts needed to replicate all results presented in this paper available after the peer review period. Elaborate details of all experiments have been provided in the Appendices.

LLM Usage Statement. We used GPT-5 solely for grammar correction and language polishing. The model was not involved in ideation, data analysis, or deriving any of the scientific contributions presented in this work.

B Dataset Construction Details

This section provides supplementary details regarding the construction of *ColorConceptBench* , including concept selection statistics, the sketch generation pipeline, and the human annotation protocol.

B.1 Concept Taxonomy & Statistics

The concept categories consist of seven classes, covering both natural and artificial objects:

- Vegetables: common vegetables, such as potatoes and bean sprouts.
- Fruit: common fruits, such as apple and banana.
- Plant: flowers and other plants, such as lily and clover.
- Animal: common animals, such as dog and bat.
- Food: processed foods and beverages, such as bagel and coffee.
- Landscape: natural scenes, such as lake and mountain.
- Building: architectural structures, both ancient and modern, such as pyramid and skyscraper.

A concept is defined either as a base word (original concept), a base word with a visual state adjective, or a base word with an emotional adjective. Visual state adjectives denote changes in the objective physical state of the concept, whereas emotional adjectives capture more abstract qualities, reflecting atmosphere or mood.

Category	Lexicon Link
Fruit & Vegetables	https://www.words-to-use.com/words/fruits-vegetables/
Water Body	https://www.yourdictionary.com/articles/water-words-descriptive-writing
Weather & Sky	https://www.writerswrite.co.za/words-to-describe-weather/
Floral	https://www.slowflowerspodcast.com/wp-content/uploads/Slow-Flowers_Creative_Workshop_Floral-Adjectives_Nouns_Verbs.pdf
Terrain	https://adjectives-for.com/terrain https://simplicable.com/descriptive-words/words-to-describe-nature
Food	https://www.words-to-use.com/words/sweets-desserts/

Table 4: Lexicon for adjectives.

Table 9 provides the detailed concept list. Each entry corresponds to an original concept and its associated visual state and emotional adjectives. Concepts are uniquely identified and grouped by category, providing a comprehensive reference for all concepts used in our experiments.

B.2 Sketch Generation Pipeline

Model Selection. To generate sketch images that are both visually clear and suitable for color annotation, we employ two complementary text-to-image generation models: Qwen-Image (Wu et al., 2025a) and Stable Diffusion 3.5 Medium (Esser et al., 2024).

Specifically, sketches generated by Qwen-Image tend to exhibit clean and simplified contours with minimal visual clutter, making them highly suitable for preserving object structure and avoiding unintended visual cues. However, such sketches may occasionally lack sufficient regions for coloring, limiting their usefulness for certain object categories. These models provide complementary strengths, allowing us to select sketches that best balance structural clarity and interior detail for color annotation.

Prompt Design. All sketches are generated using a unified prompt template designed to suppress color and texture cues while preserving structural information. Specifically, we adopt the following prompt formulation:

"A simple cartoon black outline drawing of [original concept], without coloring, without shadow, white background."

This prompt explicitly enforces the absence of color, shading, and lighting effects, ensuring that the resulting images convey only the structural characteristics of the target concept. By standardizing the prompt across all concepts, we minimize stylistic variation that could otherwise influence participants' color perception.

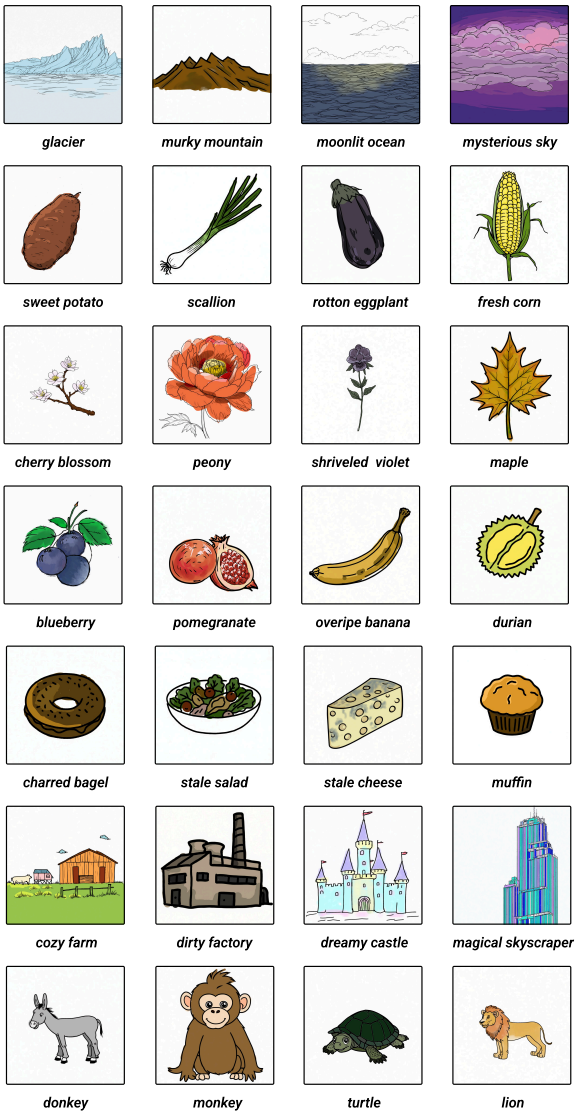

Figure 6: A gallery of our human annotated dataset.

Generation Settings and Bias Control. For the same concept, different visual state or emotional modifiers are applied using the same selected sketch, ensuring consistency across variations. With 216 original concepts in total, this results in 216 sketches. However, line drawings can sometimes contain unwanted visual hints due to complex shapes. To address this, we manually review and select the sketches. For each concept, we generate five candidates using a guidance scale of 5 to 7. From these, we choose the one with the best structural clarity and no unintended shading or texture. From the generated candidates, we

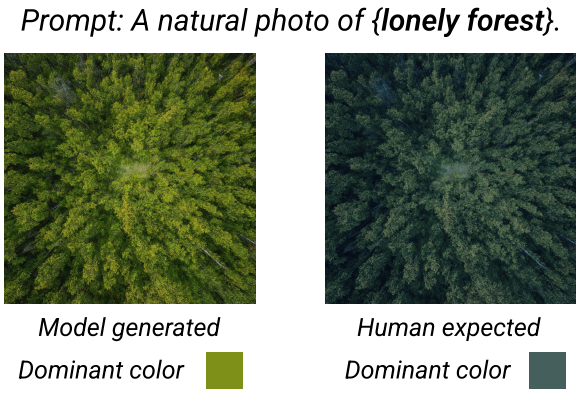

Figure 7: Misalignment color-concept association with human expectation.

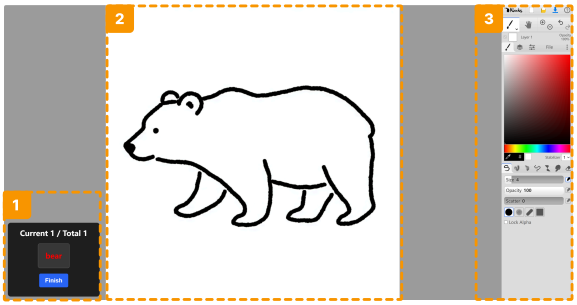

Figure 8: Annotation system interface

select one sketch that best satisfies predefined criteria. Images with artifacts, messy backgrounds, or ambiguous structures are discarded (see Figure 9 for filtered out examples). By incorporating controlled multi-sample generation and selection into the pipeline, we reduce the impact of random generation artifacts. Representative examples of the finalized sketches are provided in Figure 9.

B.3 Human Annotation Process

Demographic Information. All annotators participating in the study were professionals with a background in design or the visual arts. Their expertise spans areas such as visual design, industrial design, and fine arts (e.g., painting and illustration). Therefore, they possess strong color perception skills, ensuring that the annotations are accurate and reliable.

Designer Instructions. Annotators were provided with detailed instructions to ensure consistent and meaningful color annotations:

1) Color Selection Guidelines. Annotators were asked to use the most common colors associated with each concept according to their own memory and understanding. The colors should reflect realistic associations rather than cartoonish or childlike styles.

2) Task-specific color association. Colors must be applied strictly according to the given concept. Annotators should focus on the concept-relevant parts of the image, leaving unrelated areas uncolored. For example,for the concept “grassland”, only the grass should be colored; mountains, sky, and other elements should remain blank. For the concept “coffee”, the beverage itself should be colored, but the cup or other containers do not require coloring.

3) Reference examples. Example images were provided to help annotators understand the target coloring goals.

4) Consent Form. All annotators were required to read and sign an informed consent form before beginning the experiment, ensuring that participation was voluntary and ethically compliant.

5) Pre-experiment practice. Before beginning the main annotation task, each annotator completed a small pre-experiment of three images. These initial annotations were manually reviewed to ensure compliance with the instructions before proceeding to the formal experiment.

Payment. We pay each participant $15 for their time and effort.

B.4 Quality Control.

To ensure the reliability and semantic validity of our dataset, we implemented a three-stage quality control protocol, combining quantitative consistency checks with expert qualitative verification.

1) Qualitative Review and Iterative Verification. We first conducted a manual review process to identify annotations that may deviate from commonsense or domain-consistent interpretations of the corresponding concepts. All colorized results were examined by domain experts with experience in visual semantics and design.

For annotations considered potentially implausible or ambiguous, the corresponding participant was asked to recolor the same concept. If the two rounds of annotations exhibited consistent color patterns, the result was retained and interpreted as a stable reflection of the participant’s internal conceptual understanding, even when it differed from more frequently observed or prototypical associations. In such cases, follow-up inquiries were

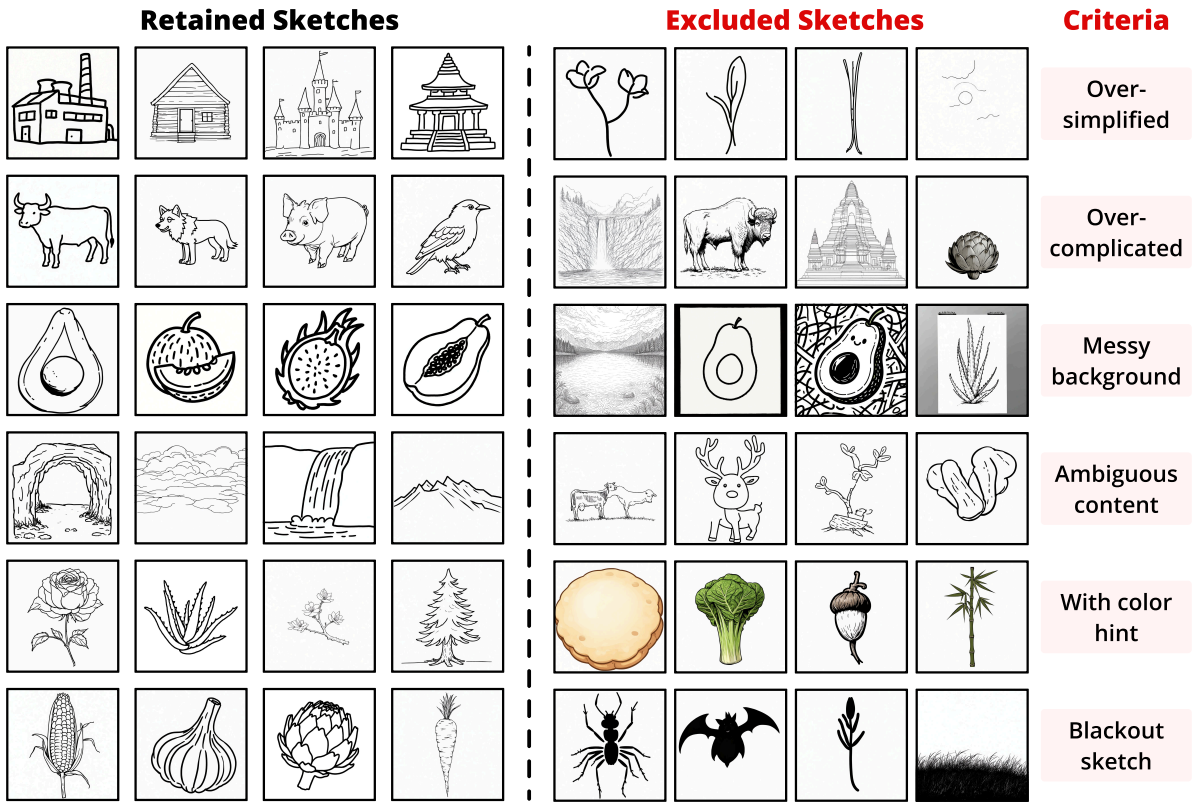

Figure 9: Our sketch

conducted to document the participant's reasoning and intended interpretation.

If the two rounds showed noticeable discrepancies, we further investigated potential causes such as misunderstanding of the target concept or misinterpretation of the provided prompt. The participant was then asked to repeat the annotation process until a stable and self-consistent result was obtained. This iterative procedure allowed us to filter out accidental or low-quality annotations and ensured that the retained data reliably reflected participants' genuine conceptual associations.

2) Quantitative Consistency Check. To guarantee high-quality ground truth, each concept was independently annotated by five professional designers who underwent specific training for this task. Specifically, for each concept c, we extracted the color distributions in the UW71 space from the five annotated images, denoted as $\{p_1, \ldots, p_5\}$. We then computed the average pairwise Earth Mover's Distance (EMD) within the group:

$$\mathrm{EMD}(c) = \frac{1}{10} \sum \mathrm{EMD}(p_i, p_j) \qquad (6)$$

where the denominator 10 represents the number of unique pairs among the five annotators. A lower score indicates higher consensus among the designers.

Expert Votes	Interpretation	Image Count
3 Yes / 0 No	Unanimous Consistent	399
0 Yes / 3 No	Unanimous Inconsistent	4
2 Yes / 1 No	Majority Consistent	127
1 Yes / 2 No	Majority Inconsistent	32
Total		**562**

Table 5: Distribution of Expert Agreement Patterns

3) Expert Verification for High-Variance Concepts. Recognizing that abstract or complex concepts may naturally exhibit higher variance (e.g., *'rotten apple'* and *'lonely cabin'*), we perform a targeted review on the "long-tail" data. We identify the top 10% of concepts with the highest average EMD scores, indicating the lowest agreement. To distinguish between valid semantic ambiguity and potential annotation errors, we recruit a separate panel of trained experts to review these high-variance samples.

- **Review Protocol:** For each concept, three independent experts review both the coloring images and color distributions.
- **Binary Validation:** The experts perform a binary classification task, labeling each copy

Figure 10: Quality control system interface

as either *Consistent* or *Inconsistent* with the semantic meaning of the concept.

- **Decision Rule:** Concepts that fail to secure a majority vote are removed from the final dataset.

Finally, we perform quality control on 562 colored images and discard 36 inconsistent samples. This hybrid approach ensures that our dataset retains rich, diverse color associations for abstract concepts while filtering out low-quality or error annotations. Samples of the verification results are shown in Figure 11.

4) Inter-Expert Agreement Analysis. To quantify inter-expert agreement during the binary validation process, we further analyze the distribution of expert votes across all reviewed images. Each image is independently evaluated by three experts and labeled as either Consistent or Inconsistent with respect to the semantic meaning of the concept.

As shown in Table 5, the majority of images receive unanimous judgments (3/3), indicating a high level of consensus among experts. A smaller portion of images exhibited partial disagreement (2 Yes / 1 No or 1 Yes / 2 No), suggesting variability in expert judgments for certain abstract concepts. Images with unanimous or majority Inconsistent votes (3 No or 2 No) were discarded from the dataset to ensure the quality and semantic consistency of the final annotations.

Based on these annotations, we compute the average observed agreement across all reviewed images. Following Fleiss' formulation, the observed agreement for each image P_i is defined as the proportion of matching expert labels, and the overall agreement is obtained by averaging over all images:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i, \tag{7}$$

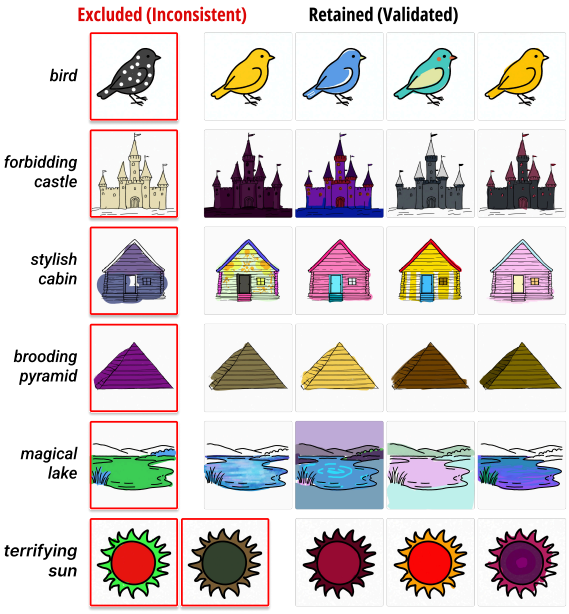

Figure 11: Examples of quality control for human-grounded color annotations. For concepts with high inter-annotator variance, we conducted a blind expert verification Y/N task. Samples marked in red (left column) were identified as semantically inconsistent outliers (e.g., a "forbidding castle" colored in bright pastels) and excluded. The retained instances (right columns) preserve the diverse but valid color distributions aligned with human cognition.

where N denotes the total number of reviewed images.

The resulting mean observed agreement is $\bar{P} = 0.811$, indicating a high level of consensus among experts despite the presence of semantically ambiguous cases.

C Implementation Details

C.1 Model Inference Configuration

To evaluate the model's capability in associating concepts with colors across different settings, we conduct a systematic image generation process. Image generation is performed using Nvidia A800 GPUs. For each concept in our benchmark, we generate images under the following protocols:

1) Visual Style. We explore *natural photo* and *clipart cartoon* as two common domains. This choice enables us to evaluate if the model captures the colors of concepts universally, or if its performance is biased towards a specific visual style.

2) Classifier-Free Guidance (CFG). We focus on the CFG scale, a hyperparameter that controls the trade-off between alignment to the text

prompt and image diversity. To investigate whether the guidance scale influences the model's color-concept association or the diversity of color selection, we evaluate the model across 7 distinct guidance scales. This aims to determine if the model's ability is robust to varying generation constraints.

3) Sampling Strategy. For each unique combination of concept, style, and guidance scale, we generate 5 independent samples at a resolution of 1024×1024 with 50 inference steps, using distinct random seeds.

C.2 Prompt Templates

We utilize standardized prompt templates to trigger concept generation across different styles:

- **Implicit Association (Ours):** "A [Style] of a [Adjective] [Object], centered composition." (e.g., *"A natural photo of a lonely street."*)

C.3 Color Extraction Pipeline

To ensure that our analysis focuses exclusively on the target concept rather than the background environment, we employ a segmentation-based extraction pipeline followed by adaptive color quantization.

Segmentation. We first use Grounding DINO (Liu et al., 2024) to identify the bounding box of the target concept. The detected bounding box is then passed to the Segment Anything Model (SAM) (Kirillov et al., 2023), which generates a binary mask of the target concept. This step ensures that the subsequent analysis is derived strictly from pixels belonging to the concept.

Quantization and Refinement. To capture the primary visual impression, we downsample the masked foreground to 100×100 pixels. We divide the RGB space into $16 \times 16 \times 16$ bins and map filtered pixels into these bins. We then implement an **adaptive grouping strategy**:

1. We calculate the CIEDE2000 distance between the dominant colors of different images, clustering images into the same visual group if their dominant colors are perceptually similar ($\Delta E_{00} \leq 12$).

2. Within each group, we aggregate pixels and quantize them using $8 \times 8 \times 8$ RGB bins.

3. The top 20 colors undergo a final merging process where color centers indistinguishable to the human eye ($\Delta E_{00} \leq 7$) are combined (Szafir, 2017).

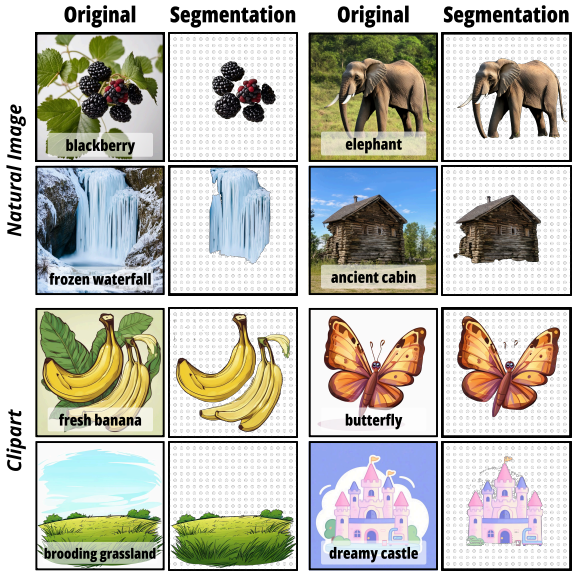

Figure 12: Color grounding using SAM.

Method	Style	
	Clipart	Natural
Flux	0.665	0.683
OmniGen2	0.904	0.734
Qwen-Image	0.868	1.029
Sana	0.722	0.627
SD 3	0.787	0.771
SD 3.5	0.864	0.767
SD XL	0.752	0.579

Table 6: Quantitative evaluation of sensitivity across different models with image style categories.

The final distribution q is constructed by aggregating these refined palettes, weighted by the number of images in each group.

D Human Judgment

We conducted a pairwise human evaluation on clipart images to validate the alignment between our metrics and human perception. We sampled three models: Sana-1.5, OmniGen2, and SDXL, representing a diverse range of EMD performance. The study utilized a Gradio interface where annotators compared two model outputs against a human ground truth. Our system employs a non-repeating random assignment mechanism to distribute tasks and randomly place the candidate models to mitigate positional bias. The study involved 36 participants (26 with design/art backgrounds and 10

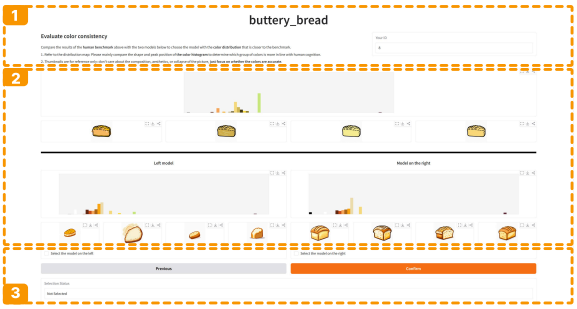

Figure 13: Human judgment system interface.

Method	Modifier	
	Visual State	Emotional
Flux	0.676	0.658
OmniGen2	0.830	0.739
Qwen-Image	0.979	0.739
Sana	0.679	0.644
SD 3	0.786	0.734
SD 3.5	0.826	0.739
SD XL	0.676	0.596

Table 7: Quantitative evaluation of sensitivity across different models with modifier categories.

non-experts). Each case received 6 independent votes, yielding a total of 2,700 valid responses.

Annotation Instructions. To ensure evaluations focused strictly on color semantics rather than image quality, we provide following instructions: (1) Participants are directed to judge primarily based on the color histogram rather than the raw image. (2) We explicitly instruct annotators to disregard composition, aesthetic appeal, or generation artifacts. The generated images are provided solely for reference, with the core task defined as "selecting the model whose color distribution best matches the human ground-truth."

Method	Style		
	Clipart	Natural	Average
Emotional	0.722	0.663	0.693
Visual State	0.805	0.753	0.779

Table 8: Quantitative evaluation of sensitivity across different categories.

E Additional Quantitative Results

We present the quantitative evaluation of sensitivity in Table 6, Table 7, and Table 8. Table 6 and Table 7 details the performance of individual models across different styles and different modifiers. To provide a broader perspective, Table 8 summarizes the results by semantic modifier category (Emotional vs. Visual State), including the calculated average sensitivity scores to facilitate a overall comparison.

F Additional Qualitative Results

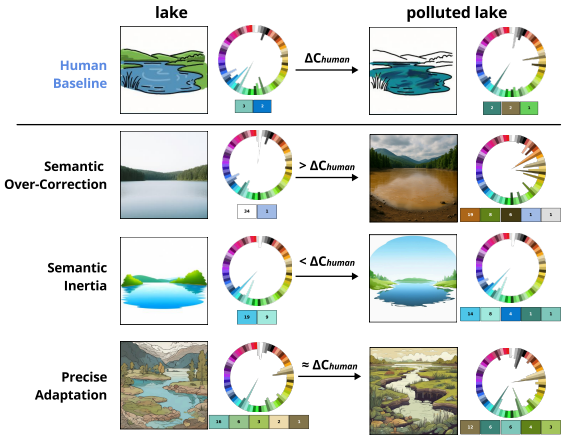

Figure 14: Qualitative results. Compared to the Human baseline (Top), models exhibit three distinct behaviors: Over-Correction (excessive shift, loss of identity), Inertia (negligible shift, failure to adapt), and Precise Adaptation (balanced shift, preserving object identity while updating attributes).

Table 9: Concept list

Id	Category	Concept	Visual state modifier	Emotional modifier
1	animal	dog		
2	animal	bird		
3	animal	horse		
4	animal	chicken		
5	animal	bear		
6	animal	cat		
7	animal	wolf		
8	animal	deer		
9	animal	bull		
10	animal	cow		
11	animal	lion		
12	animal	duck		
13	animal	pig		
14	animal	bat		
15	animal	whale		
16	animal	seal		
17	animal	bee		
18	animal	sheep		
19	animal	elephant		
20	animal	shark		
21	animal	rabbit		
22	animal	monkey		
23	animal	goat		
24	animal	butterfly		
25	animal	crab		
26	animal	frog		
27	animal	turtle		
28	animal	crow		
29	animal	goose		
30	animal	spider		
31	animal	ant		
32	animal	dolphin		
33	animal	lobster		
34	animal	owl		
35	animal	coral		
36	animal	squirrel		
37	animal	camel		
38	animal	pigeon		
39	animal	swan		
40	animal	donkey		
41	fruit	apple	fresh, unripe, overripe, rotten, juiceless	
42	fruit	apricot	fresh, unripe, overripe, rotten, juiceless	
43	fruit	avocado	fresh, unripe, overripe, rotten, juiceless	
44	fruit	banana	fresh, unripe, overripe, rotten, juiceless	
45	fruit	blackberry	fresh, unripe, overripe, rotten, juiceless	
46	fruit	blueberry	fresh, unripe, overripe, rotten, juiceless	
47	fruit	cantaloupe	fresh, unripe, overripe, rotten, juiceless	
48	fruit	cherry	fresh, unripe, overripe, rotten, juiceless	
49	fruit	coconut	fresh, unripe, overripe, rotten, juiceless	
50	fruit	cranberry	fresh, unripe, overripe, rotten, juiceless	
51	fruit	dragonfruit	fresh, unripe, overripe, rotten, juiceless	
52	fruit	durian	fresh, unripe, overripe, rotten, juiceless	
53	fruit	fig	fresh, unripe, overripe, rotten, juiceless	
54	fruit	grape	fresh, unripe, overripe, rotten, juiceless	
55	fruit	grapefruit	fresh, unripe, overripe, rotten, juiceless	
56	fruit	kiwi	fresh, unripe, overripe, rotten, juiceless	
57	fruit	lemon	fresh, unripe, overripe, rotten, juiceless	
58	fruit	lime	fresh, unripe, overripe, rotten, juiceless	
59	fruit	lychee	fresh, unripe, overripe, rotten, juiceless	
60	fruit	mango	fresh, unripe, overripe, rotten, juiceless	
61	fruit	melon	fresh, unripe, overripe, rotten, juiceless	
62	fruit	mulberry	fresh, unripe, overripe, rotten, juiceless	
63	fruit	olive	fresh, unripe, overripe, rotten, juiceless	
64	fruit	orange	fresh, unripe, overripe, rotten, juiceless	
65	fruit	papaya	fresh, unripe, overripe, rotten, juiceless	
66	fruit	peach	fresh, unripe, overripe, rotten, juiceless	
67	fruit	pear	fresh, unripe, overripe, rotten, juiceless	
68	fruit	pineapple	fresh, unripe, overripe, rotten, juiceless	
69	fruit	plum	fresh, unripe, overripe, rotten, juiceless	
70	fruit	pomegranate	fresh, unripe, overripe, rotten, juiceless	
71	fruit	pomelo	fresh, unripe, overripe, rotten, juiceless	
72	fruit	raspberry	fresh, unripe, overripe, rotten, juiceless	
73	fruit	star fruit	fresh, unripe, overripe, rotten, juiceless	
74	fruit	strawberry	fresh, unripe, overripe, rotten, juiceless	
75	fruit	watermelon	fresh, unripe, overripe, rotten, juiceless	
76	food	bagel	browned, buttery, charred, chocolaty, cinnamon, nutty, stale	
77	food	biscuit	browned, buttery, charred, chocolaty, cinnamon, nutty, stale	
78	food	bread	browned, buttery, charred, chocolaty, cinnamon, nutty, stale, toasted	
79	food	cookie	browned, buttery, charred, chocolaty, cinnamon, cooled, nutty, stale	
80	food	nut	browned, buttery, charred, cinnamon, stale, toasted	
81	food	pancake	browned, buttery, charred, chocolaty, cinnamon, nutty, stale, vanilla	

Id	Category	Concept	Visual state modifier	Emotional modifier
82	food	sandwich	browned, charred, chocolaty, cinnamon, stale	
83	food	toast	browned, buttery, charred, chocolaty, cinnamon, nutty, stale	
84	food	beer	buttery, cooled, stale	
85	food	brownie	buttery, stale	
86	food	cereal	buttery, charred, chocolaty, stale	
87	food	dumpling	buttery, charred, stale	
88	food	hamburger	buttery, charred, stale	
89	food	muffin	buttery, chocolaty, nutty, stale, vanilla	
90	food	oatmeal	buttery, chocolaty, cinnamon, nutty, stale, vanilla	
91	food	pie	buttery, charred, chocolaty, cinnamon, stale	
92	food	pudding	buttery, chocolaty, cinnamon, nutty, stale, vanilla	
93	food	curry	charred, stale	
94	food	egg	charred, stale, toasted	
95	food	pasta	charred, stale	
96	food	pizza	charred, stale	
97	food	steak	charred, stale	
98	food	butter	chocolaty, stale	
99	food	candy	chocolaty, nutty, stale	
100	food	milk	chocolaty, cinnamon, nutty, stale, vanilla	
101	food	yogurt	chocolaty, cinnamon, cooled, nutty, stale, vanilla	
102	food	coffee	cinnamon, cooled, nutty, stale, vanilla	
103	food	tea	cinnamon, stale	
104	food	jelly	cooled, stale	
105	food	juice	cooled, stale	
106	food	water	cooled	
107	food	chocolate	nutty, stale, vanilla	
108	food	birthday cake	stale	
109	food	cheese	stale	
110	food	honey	stale	
111	food	rice	stale	
112	food	salad	stale	
113	food	sushi	stale	
114	plant	acorn	shrivelled, fresh	
115	plant	aloe	shrivelled, fresh	
116	plant	bamboo	shrivelled, fresh	
117	plant	basil	shrivelled, fresh	
118	plant	bush	shrivelled, fresh	
119	plant	cactus	shrivelled, fresh	
120	plant	carnation	shrivelled, fresh	
121	plant	cherry blossom	shrivelled, fresh	
122	plant	clover	shrivelled, fresh	
123	plant	daisy	shrivelled, fresh	
124	plant	dandelion	shrivelled, fresh	
125	plant	ginkgo	shrivelled, fresh	
126	plant	ivy	shrivelled, fresh	
127	plant	jasmine	shrivelled, fresh	
128	plant	lavender	shrivelled, fresh	
129	plant	lily	shrivelled, fresh	
130	plant	lotus	shrivelled, fresh	
131	plant	maple	shrivelled, fresh	
132	plant	marigold	shrivelled, fresh	
133	plant	mint	shrivelled, fresh	
134	plant	moss	shrivelled, fresh	
135	plant	oak	shrivelled, fresh	
136	plant	orchid	shrivelled, fresh	
137	plant	palm tree	shrivelled, fresh	
138	plant	peony	shrivelled, fresh	
139	plant	pine tree	shrivelled, fresh	
140	plant	rose	shrivelled, fresh	
141	plant	seaweed	shrivelled, fresh	
142	plant	straw	shrivelled, fresh	
143	plant	sunflower	shrivelled, fresh	
144	plant	tulip	shrivelled, fresh	
145	plant	violet	shrivelled, fresh	
146	plant	wheat	shrivelled, fresh	
147	vegetables	artichoke	fresh, shrivelled, rotten, juiceless	
148	vegetables	arugula	fresh, shrivelled, rotten, juiceless	
149	vegetables	asparagus	fresh, shrivelled, rotten, juiceless	
150	vegetables	bean	fresh, shrivelled, rotten, juiceless	
151	vegetables	bell pepper	fresh, shrivelled, rotten, juiceless	
152	vegetables	bok choy	fresh, shrivelled, rotten, juiceless	
153	vegetables	broccoli	fresh, shrivelled, rotten, juiceless	
154	vegetables	brussels sprouts	fresh, shrivelled, rotten, juiceless	
155	vegetables	cabbage	fresh, shrivelled, rotten, juiceless	
156	vegetables	carrot	fresh, shrivelled, rotten, juiceless	
157	vegetables	cauliflower	fresh, shrivelled, rotten, juiceless	
158	vegetables	celery	fresh, shrivelled, rotten, juiceless	
159	vegetables	chive	fresh, shrivelled, rotten, juiceless	
160	vegetables	corn	fresh, shrivelled, rotten, juiceless	
161	vegetables	cucumber	fresh, shrivelled, rotten, juiceless	
162	vegetables	eggplant	fresh, shrivelled, rotten, juiceless	
163	vegetables	garlic	fresh, shrivelled, rotten, juiceless	
164	vegetables	gourd	fresh, shrivelled, rotten, juiceless	
165	vegetables	jalapeno	fresh, shrivelled, rotten, juiceless	
166	vegetables	kale	fresh, shrivelled, rotten, juiceless	
167	vegetables	leek	fresh, shrivelled, rotten, juiceless	

Id	Category	Concept	Visual state modifier	Emotional modifier
168	vegetables	lettuce	fresh, shrivelled, rotten, juiceless	
169	vegetables	okra	fresh, shrivelled, rotten, juiceless	
170	vegetables	onion	fresh, shrivelled, rotten, juiceless	
171	vegetables	parsley	fresh, shrivelled, rotten, juiceless	
172	vegetables	pea	fresh, shrivelled, rotten, juiceless	
173	vegetables	pepper	fresh, shrivelled, rotten, juiceless	
174	vegetables	pickle	fresh, shrivelled, rotten, juiceless	
175	vegetables	potato	fresh, shrivelled, rotten, juiceless	
176	vegetables	pumpkin	fresh, shrivelled, rotten, juiceless	
177	vegetables	radish	fresh, shrivelled, rotten, juiceless	
178	vegetables	scallion	fresh, shrivelled, rotten, juiceless	
179	vegetables	spinach	fresh, shrivelled, rotten, juiceless	
180	vegetables	sprouts	fresh, shrivelled, rotten, juiceless	
181	vegetables	squash	fresh, shrivelled, rotten, juiceless	
182	vegetables	sweet potato	fresh, shrivelled, rotten, juiceless	
183	vegetables	tomato	fresh, shrivelled, rotten, juiceless	
184	vegetables	zucchini	fresh, shrivelled, rotten, juiceless	
185	landscape	beach	arid, autumnal, brooding, desolate, dirty, fiery, magical, misty, moonlit, polluted, rainy, spring, stormy, summery, sunny, tropical, vibrant	lonely, oppressive, serene, terrifying, warm
186	landscape	cave	arid, autumnal, brooding, desolate, dirty, forbidding, grassy, magical, misty, muddy, mysterious	lonely, oppressive, serene, somber, terrifying
187	landscape	desert	arid, barren, brooding, desolate, fiery, forbidding, magical, misty, moonlit, mysterious, polluted, rainy, spectacular, spring, summery	oppressive, serene, terrifying
188	landscape	forest	arid, autumnal, barren, brooding, dense, desolate, ethereal, fiery, forbidding, lush, magical, misty, moonlit, muddy, mysterious, rainy, spring, stormy, summery, sunny, tropical, verdant, vibrant	cozy, lonely, oppressive, serene, terrifying, warm
189	landscape	grassland	arid, autumnal, barren, brooding, desolate, forbidding, magical, misty, moonlit, muddy, mysterious, olive, rainy, spring, stormy, summery, sunny, tropical, verdant, vibrant	cozy, lonely, oppressive, serene, warm
190	landscape	lake	arid, autumnal, blushing, brooding, clear, dirty, ethereal, forbidding, frozen, magical, misty, moonlit, murky, mysterious, polluted, rainy, spring, stormy, summery, sunny, tropical, turbulent, vibrant	lonely, oppressive, serene, terrifying, warm
191	landscape	mountain	arid, autumnal, barren, brooding, desolate, forbidding, grassy, magical, majestic, misty, moonlit, muddy, mysterious, olive, rainy, spectacular, spring, stormy, summery, sunny, tropical, verdant, vibrant	lonely, oppressive, serene, terrifying
192	landscape	island	autumnal, barren, brooding, desolate, dirty, dreamy, forbidding, grassy, magical, misty, moonlit, muddy, mysterious, polluted, rainy, spring, stormy, summery, tropical, vibrant	lonely, oppressive, serene, terrifying
193	landscape	sunrise	autumnal, blushing, fiery, magical, misty, spectacular, spring, summery, tropical	cozy, hopeful, oppressive, serene, terrifying, warm
194	landscape	sunset	autumnal, blushing, fiery, magical, misty, spectacular, spring, summery, tropical	cozy, oppressive, serene, terrifying, warm
195	landscape	moon	blushing, brooding, forbidding, magical, mysterious, pale, summery	oppressive, somber, terrifying, wistful
196	landscape	sky	blushing, brooding, clear, cloudless, cloudy, faded, fiery, forbidding, magical, misty, murky, mysterious, pale, polluted, rainy, spring, starry, stormy, sunny, thunderous, tropical	cozy, hopeful, oppressive, serene, somber, terrifying, wistful
197	landscape	sun	blushing, fiery, magical, summery, tropical	hopeful, oppressive, terrifying, warm
198	landscape	glacier	brooding, forbidding, magical, majestic, misty, mysterious, polluted, rainy, spectacular, spring	oppressive, serene
199	landscape	ocean	brooding, clear, desolate, dirty, forbidding, frozen, magical, misty, moonlit, murky, mysterious, polluted, rainy, spectacular, spring, stormy, summery, sunny, tropical, turbulent	oppressive, serene, terrifying, warm
200	landscape	waterfall	brooding, dirty, ethereal, forbidding, frozen, magical, majestic, misty, moonlit, murky, mysterious, polluted, spectacular, spring, stormy, turbulent	oppressive, terrifying
201	landscape	rainbow	magical	hopeful
202	building	cabin	ancient, autumnal, brooding, dirty, forbidding, grassy, magical, mysterious, stormy, stylish, vintage	somber, terrifying, warm, oppressive, serene, cozy, lonely
203	building	castle	ancient, autumnal, brooding, desolate, dirty, dreamy, forbidding, grassy, magical, majestic, mysterious, solemn, spectacular, stormy, stylish, sumptuous, vintage	somber, terrifying, oppressive, serene, lonely
204	building	church	ancient, autumnal, brooding, desolate, dirty, dreamy, forbidding, grassy, magical, majestic, mysterious, solemn, spectacular, stormy, stylish, sumptuous, vintage	somber, terrifying, oppressive, serene, lonely
205	building	hospital	ancient, brooding, dirty, forbidding, magical, mysterious	somber, terrifying, oppressive, serene
206	building	palace	ancient, autumnal, brooding, desolate, dirty, dreamy, forbidding, grassy, magical, majestic, mysterious, solemn, spectacular, stormy, stylish, sumptuous, vintage	somber, terrifying, oppressive, serene, lonely
207	building	pyramid	ancient, autumnal, brooding, desolate, forbidding, magical, majestic, mysterious, solemn, spectacular, stormy	somber, terrifying, oppressive, serene, lonely
208	building	temple	ancient, autumnal, brooding, desolate, dirty, fiery, forbidding, grassy, magical, majestic, mysterious, solemn, spectacular, stormy	somber, terrifying, oppressive, serene, lonely
209	building	farm	arid, autumnal, barren, brooding, desolate, dirty, forbidding, magical, misty, mysterious, rainy, spring, stormy, summery, tropical, verdant, vibrant	hopeful, terrifying, serene, oppressive, cozy
210	building	aquarium	brooding, dirty, dreamy, forbidding, magical, mysterious, spectacular, sumptuous	somber, oppressive, serene
211	building	casino	brooding, dreamy, forbidding, magical, mysterious, spectacular, sumptuous, vintage	somber, oppressive, serene

Id	Category	Concept	Visual state modifier	Emotional modifier
212	building	factory	brooding, desolate, dirty, forbidding, magical, mysterious, spectacular, vintage	somber, terrifying, oppressive, serene, lonely
213	building	igloo	brooding, desolate, forbidding, magical, mysterious	somber, oppressive, serene, lonely
214	building	skyscraper	brooding, forbidding, magical, majestic, mysterious, spectacular, stylish, sumptuous	somber, terrifying, oppressive, serene, lonely