Supplementary: Prompt-Driven Color Accessibility Evaluation in Diffusion-based Image Generation Models

Xinyao Zhuang
University College London
London, United Kingdom
xinyao.zhuang.24@ucl.ac.uk

Jose Echevarria
Adobe Research
New York, USA
echevarr@adobe.com

Kaan Akşit
University College London
London, United Kingdom
k.aksit@ucl.ac.uk

1 Prompt Design for Dataset Construction

Table 1 lists the **standard** prompts used for each content category. All other prompt variants (**colorblind-aware**, **protanopia-aware**, and **deuteranopia-aware**) were generated by appending consistent accessibility-related modifiers to these base prompts. For clarity and conciseness, only the **standard** prompts are reported here.

Category	Standard Prompt
Candy	A plate full of rainbow candy.
Cartoon	A colorful cartoon scene.
Coral Reef	A vivid coral reef underwater.
Flower	A vibrant bouquet of mixed flowers.
Fruit	A bowl of fruit.
Parrot	Colorful parrots perched on a branch.
Poster	A vivid musical festival poster.
Street View	A colorful bustling urban city street view.

Table 1: Standard prompts used for each content category.

Figure 1 shows representative images from the eight categories used in our dataset. These examples illustrate the visual diversity of the evaluation set, including natural scenes, synthetic graphics, object-centered images, and urban scenes.

2 Experimental Settings for Generation Models

For the main evaluation, we generated images using two pretrained text-to-image models: Stable Diffusion 3.5-large [StabilityAI [n. d.]] and FLUX.2 Klein 4B [Labs 2026]. For both models, we used the same prompt set, the same eight categories, and the same number of generated samples per prompt. All images used for *CVDLoss* evaluation were generated at 1024×1024 resolution.

Stable Diffusion 3.5-large was run on an NVIDIA RTX 4090 GPU with 4-bit NF4 quantization, Bfloat16 compute precision, and CPU offloading of model components. Images were generated with 28 inference steps and a guidance scale of 4.5.

FLUX.2 Klein 4B was run using the Flux2KleinPipeline in Bfloat16 precision on CUDA. We used the model-specific recommended sampling setting of 4 inference steps. Since FLUX.2 Klein is a step-wise distilled model, classifier-free guidance scale is not an active control parameter. Therefore, we fixed the prompt set, image resolution, number of samples, and *CVDLoss* evaluation pipeline across models while using model-specific inference settings.

3 Color Vision Deficiencies Simulation

We simulated Color Vision Deficiencies(CVD) using the simulate module of the DaltonLens library [Burrus 2021], which includes the physiologically based simulation model proposed by Machado et al. [Machado et al. 2009]. For all simulations, the severity parameter was set to 1.0, corresponding to complete protanopia (red-blindness) or deuteranopia (green-blindness). In cases where prompts were specifically designed for one type of deficiency (e.g., **protanopia-aware** prompts), only the corresponding simulation was performed. For more general prompts (**standard** and **colorblind-aware**), both protanopia and deuteranopia simulations were applied and analyzed.

Figure 1: Examples from the eight categories in our dataset.

4 Color Accessibility Measurement

4.1 Color-Space Conversion

Both original and CVD-simulated images were transformed from sRGB into the perceptually uniform OKLab color space using the colour-science Python library [Developers 2025]. The conversion pipeline followed:

(1) **Normalization**: sRGB images were normalized to floating point $[0, 1]$ arrays.
(2) **sRGB to XYZ**: sRGB was transformed to CIE XYZ using sRGB primaries and D65 whitepoint.
(3) **XYZ to OKLab**: OKLab coordinates were computed to enable perceptual gradient analysis.

4.2 Color Gradient Magnitude Map and CVDLoss

To capture structural and chromatic differences, we employed a HyAB-based gradient magnitude metric in OKLab space. For two color vectors $a = (L_a, A_a, B_a)$ and $b = (L_b, A_b, B_b)$, the HyAB difference is defined as:

$$\Delta E_{\text{HyAB}}(a, b) = |L_a - L_b| + \sqrt{(A_a - A_b)^2 + (B_a - B_b)^2}. \quad (1)$$

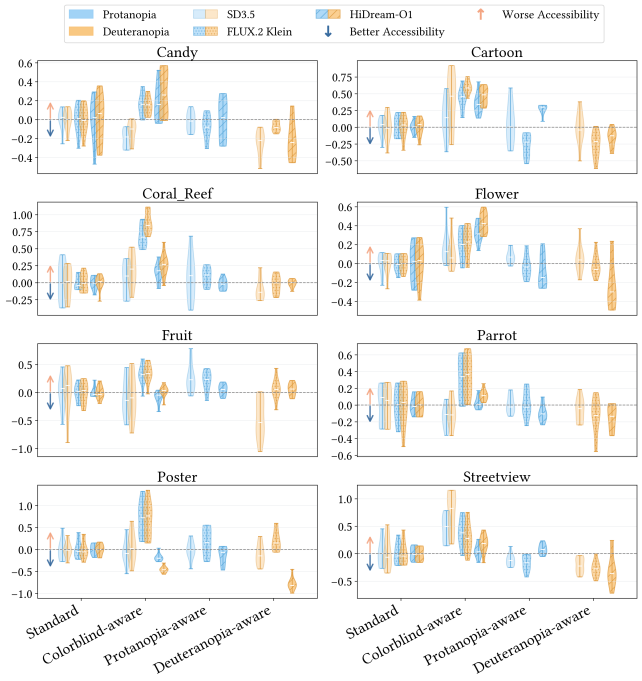

Figure 2: Additional normalized *CVDLoss* comparison across Stable Diffusion 3.5-large, FLUX.2 Klein, and HiDream-O1. Values are shown in log scale and normalized by subtracting the mean $\log_{10}(\text{CVDLoss})$ of the standard prompt for each model, category, and CVD type. Negative values indicate reduced perceptual-structural discrepancy under CVD simulation, while positive values indicate worse accessibility.

Gradient magnitude maps were then computed using Scharr operators, defined as weighted differences of HyAB values across neighboring pixels. For horizontal (G_x) and vertical (G_y) directions:

$$G_x(i,j) = w_0 \cdot \Delta E(z_1, z_7) + w_1 \cdot \Delta E(z_2, z_8) + w_0 \cdot \Delta E(z_3, z_9), \quad (2)$$

$$G_y(i,j) = w_0 \cdot \Delta E(z_1, z_3) + w_1 \cdot \Delta E(z_4, z_6) + w_0 \cdot \Delta E(z_7, z_9), \quad (3)$$

where $z_1 \dots z_9$ denote the 3×3 neighborhood of a pixel, and $w_0 = 47.0$, $w_1 = 162.0$ are Scharr operators. The overall gradient magnitude at pixel (i, j) is:

$$G(i,j) = \sqrt{G_x(i,j)^2 + G_y(i,j)^2}. \quad (4)$$

Finally, we define our *CVDLoss* as the mean squared error between the gradient magnitude maps of the original image (G_{orig}) and its CVD-simulated counterpart (G_{cvd}), normalized by the squared maximum gradient magnitude of the original image:

$$\text{CVDLoss} = \frac{\sum_p \left(G_{\text{orig}}(p) - G_{\text{cvd}}(p)\right)^2}{N \cdot \max_p \left(G_{\text{orig}}(p)\right)^2}, \quad (5)$$

where (i, j) denotes spatial pixel coordinates, while p indexes pixels using a linearized representation of the image, and N is the total number of pixels.

5 Additional Results with HiDream-O1

We used HiDream-O1-Image-Dev-2604 [HiDream-ai 2026] as an additional text-to-image model. The model was run on an NVIDIA RTX 4090 GPU with the official HiDream-O1 inference code. For consistency with the main experiments, we used the same prompt set, eight categories, four prompt types, and 10 samples per prompt. All images were generated at 1024×1024 resolution and evaluated using the same DaltonLens-based CVD simulation and *CVDLoss* pipeline.

Figure 2 shows the normalized *CVDLoss* distributions when HiDream-O1 is added to the comparison. The additional model supports the same qualitative conclusion observed in the main manuscript: accessibility-oriented prompts do not consistently reduce *CVDLoss* across categories or CVD types. In several categories, the general **colorblind-aware** prompt remains above the **standard** prompt baseline, indicating worse perceptual-structural preservation under simulated CVD. Deficiency-specific prompts sometimes reduce *CVDLoss*, especially for **deuteranopia-aware** prompts in categories such as *candy*, *parrot*, and *street view*, but the effect remains category-dependent rather than universally beneficial.

References

Nicolas Burrus. 2021. DaltonLens Python. https://github.com/DaltonLens/DaltonLens-Python. Accessed 07-08-2025.

Colour Developers. 2025. Colour Science for Python. https://pypi.org/project/colour-science/. [Accessed 22-10-2025].

HiDream-ai. 2026. HiDream-O1-Image. https://huggingface.co/HiDream-ai/HiDream-O1-Image. [Accessed 10-07-2026].

Black Forest Labs. 2026. FLUX.2 Klein 4B. https://huggingface.co/black-forest-labs/FLUX.2-klein-4B. [Accessed 30-06-2026].

G. M. Machado, M. M. Oliveira, and L. A. F. Fernandes. 2009. A Physiologically-based Model for Simulation of Color Vision Deficiency. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (November/December 2009), 1291–1298.

StabilityAI. [n. d.]. Stable Diffusion 3.5 Large. https://huggingface.co/stabilityai/stable-diffusion-3.5-large. [Accessed 13-08-2025].