# Color Conditional Generation with Sliced Wasserstein Guidance

Alexander Lobashev† Maria Larchenko* Dmitry Guskov†

† Glam AI, San Francisco, USA * Magicly AI, Dubai, UAE

## Overview & Motivation

**Problem.** Natural-language prompts are not expressive enough to reliably control colors in text-to-image diffusion models. Post-hoc color transfer methods can distort semantics and introduce style leakage.

**Idea.** Guide the denoising trajectory *during sampling* using a differentiable distance between the *color distributions* of the current prediction and a reference palette.

**Method. SW-Guidance** adds a training-free guidance term based on the *Sliced 1-Wasserstein (SW)* distance between RGB color distributions. It optimizes a small control variable $u$ at each scheduler step to align palettes while preserving prompt semantics.

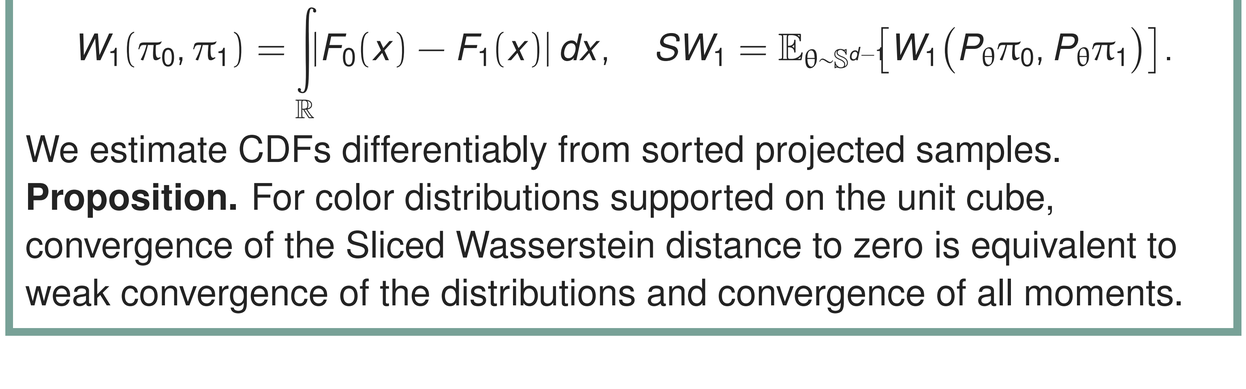
**Contributions.**

- First use of differentiable **Sliced 1-Wasserstein** color loss inside diffusion sampling for palette control.
- **State-of-the-art** palette matching on SDXL, with strong text alignment and visual quality.
- Simple, training-free plug-in; compatible with ControlNets and classifier-free guidance.

## Qualitative Palette Control

SW-Guidance achieves a strong palette match *without* transferring unwanted high-level style or textures.

## Method: SW-Guidance Inside Sampling




General scheme for SW-Guidance with latent diffusion (decoder $D$) and color-feature extractor $\Psi$ (RGB pixels).

Let $x_T \sim \mathcal{N}(0, I)$, and $x_t$ be the latent at step $t$. We optimize a small control $u$ per step:

$$x_t' = x_t + u, \qquad \hat{x}_0(x_t') \to \hat{y}_0 = D(\hat{x}_0), \qquad \mathcal{L}(u) = SW_1\left(\pi_{\hat{y}_0(u)}, \pi_{\text{ref}}\right).$$

We perform $M$ inner gradient steps on $u$ and then apply the DDIM update with classifier-free guidance.

**Color loss (1D closed form).** For 1D projections the $W_1$ distance equals the $L^1$ gap of CDFs:

$$W_1(\pi_0, \pi_1) = \int_{\mathbb{R}} |F_0(x) - F_1(x)|\, dx, \quad SW_1 = \mathbb{E}_{\theta \sim \mathbb{S}^d}\left[W_1\left(P_\theta \pi_0, P_\theta \pi_1\right)\right].$$

We estimate CDFs differentiably from sorted projected samples.
**Proposition.** For color distributions supported on the unit cube, convergence of the Sliced Wasserstein distance to zero is equivalent to weak convergence of the distributions and convergence of all moments.

## Main Results on SDXL (RealVisXL/DS, DDIM-30)

**Test setting.** 1000 ContraStyles prompts × 1000 Unsplash-Lite palettes. Metrics: **W2** between RGB color distributions (lower is better), **CLIP-IQA** (visual quality), **CLIP-T** (text-image alignment).

Table: Palette similarity (↓ W2).

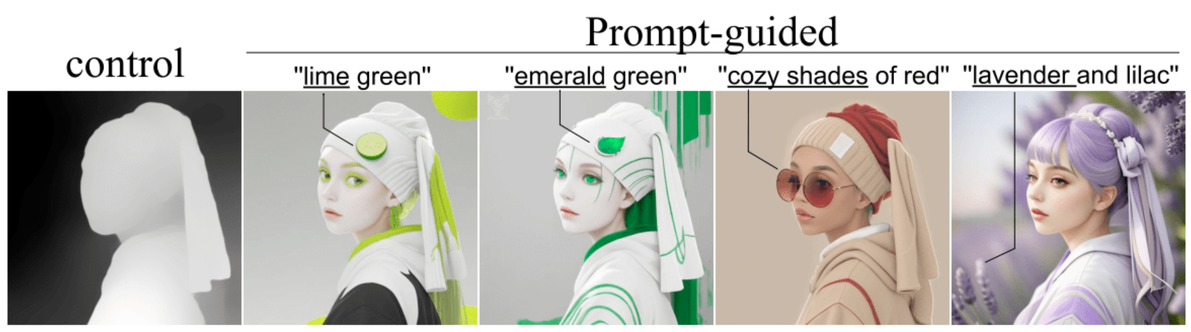
| Method (SDXL) | W2 |
|---|---|
| SW-Guidance (ours) | **0.0297 ± 0.0005** |
| hm-mkl-hm | 0.0543 ± 0.0011 |
| hm | 0.0856 ± 0.0016 |
| PhotoWCT2 | 0.1028 ± 0.0014 |
| ModFlows | 0.1125 ± 0.0016 |
| MKL | 0.1191 ± 0.0017 |
| CT | 0.1333 ± 0.0018 |
| WCT2 | 0.1347 ± 0.0017 |
| PhotoNAS | 0.1608 ± 0.0017 |
| InstantStyle | 0.1758 ± 0.0028 |
| IP-Adapter | 0.2193 ± 0.0032 |
| Unconditional SDXL | 0.3824 ± 0.0059 |

Table: Content & quality (CLIP ↑).

| Method (SDXL) | CLIP-IQA | CLIP-T |
|---|---|---|
| SW-Guidance (ours) | 0.285 ± 0.004 | 0.270 ± 0.002 |
| hm-mkl-hm | 0.259 ± 0.003 | 0.277 ± 0.002 |
| hm | 0.244 ± 0.003 | 0.282 ± 0.002 |
| PhotoWCT2 | 0.225 ± 0.003 | 0.276 ± 0.002 |
| ModFlows | 0.257 ± 0.003 | 0.282 ± 0.002 |
| MKL | 0.238 ± 0.003 | 0.283 ± 0.002 |
| CT | 0.230 ± 0.003 | 0.284 ± 0.002 |
| WCT2 | 0.179 ± 0.002 | 0.288 ± 0.002 |
| PhotoNAS | 0.167 ± 0.002 | 0.279 ± 0.002 |
| InstantStyle | **0.332 ± 0.003** | 0.238 ± 0.002 |
| IP-Adapter | 0.247 ± 0.002 | 0.214 ± 0.002 |
| Unconditional SDXL | 0.239 ± 0.003 | **0.294 ± 0.002** |

**Takeaways.** SW-Guidance has the **best palette match** and the **best prompt adherence among stylized methods**, with strong visual quality.

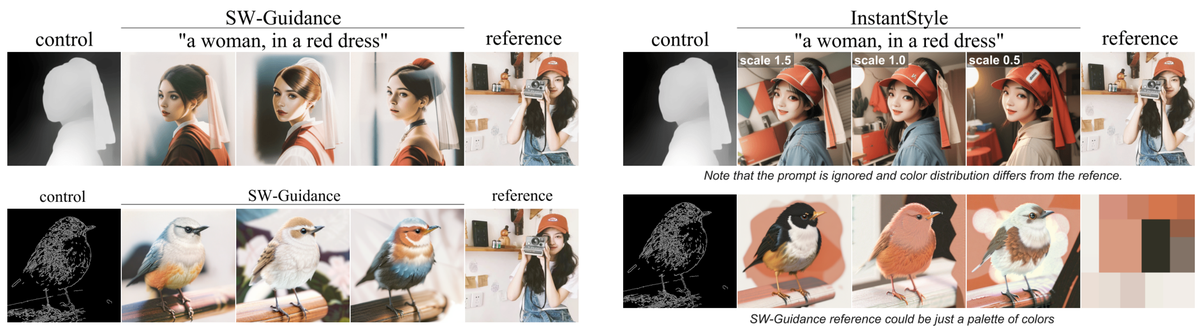
## Why Text Prompts Alone Are Not Enough for Color Control




Text description of a color may introduce unwanted content details.

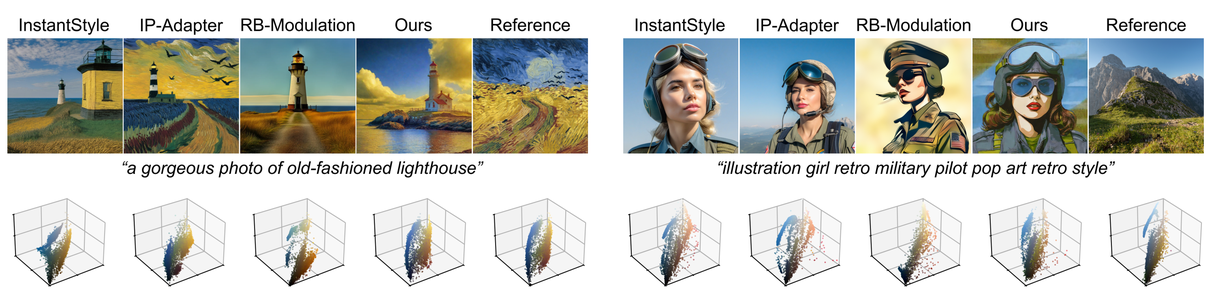
## Compatibility with Other Controls

**Works with ControlNets.** SW-Guidance composes well with edge/depth layout controls while keeping color disjoint from high-level style.




Depth/Canny control + SW-Guidance (SD1.5). Palette control remains precise and disentangled.

**Stylization vs Color.** Attention-based stylizers (e.g., IP-Adapter/InstantStyle) often entangle color with textures/composition. SW-Guidance targets *only* the RGB palette.




Compared to stylizers, SW-Guidance achieves tighter palette matching without importing unwanted semantic/style patterns.

## Limitations & Practical Notes

**Prompt collisions.** A prompt which specifies colors contradicting the reference palette can cause artifacts.
**Runtime.** With DDIM-30, $K = 10$ slices and $M = 10$ inner steps, SD-1.5 ~30s, SDXL ~1min on an RTX 4090.

## Code & Contact

**Alexander Lobashev**
lobashevalexander@gmail.com

**Maria Larchenko**
mariia.larchenko@gmail.com

**Dmitry Guskov**
guskov01dmitry@gmail.com

**GitHub:**
`alobashev/sw-guidance`