CONTROLLING STYLE IN DIFFUSION MODELS THROUGH NOISE

Everaert M.N., Süsstrunk S., Achanta R., EPFL
{martin.everaert, sabine.susstrunk, radhakrishna.achanta}@epfl.ch

ICVSS 2024
Sicily ~ 7-13 July 2024
International Computer Vision Summer School

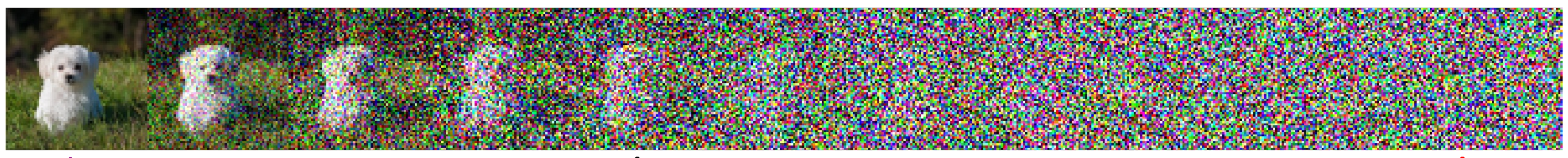
Abstract

We observe that the style of images generated by Stable Diffusion is tied to the initial noise. Thus, we propose a method to adapt Stable Diffusion to various styles using style-specific noise during fine-tuning (ICCV23). We subsequently explain that white noise added during training preserves low-frequency (LF) content, and the model then learns to maintain the LF of the initial noise. Controlling this initial noise allows to generate images with desired styles without fine-tuning (WACV24).

Diffusion models

Forward Diffusion (Noising)

x_0 $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon$ with $\varepsilon \sim \mathcal{N}(0_d, \mathrm{I}_{d\times d})$ x_{1000}



$\hat{x}_0$ $\hat{x}_t$ $\hat{x}_{1000}$

Reverse Diffusion (Generation/Sampling)

Image from [2]

Diffusion in Style

Everaert M.N., Bocchio M., Arpa S., Süsstrunk S., Achanta R. — ICCV23 PARIS

The initial noise $\hat{x}_{1000}$ affects the style of the generated image $\hat{x}_0$, so adapting it to the style facilitates style adaptation.

We fine-tune Stable Diffusion (SD) [1] with a **style-specific noise distribution** $\mathcal{N}(\mu_{style}, \Sigma_{style})$ instead of the default $\mathcal{N}(0_d, \mathrm{I}_{d\times d})$.

Original diffusion
$\varepsilon \sim \mathcal{N}(0_d, \mathrm{I}_{d\times d})$

Our style-adapted diffusion
$\varepsilon \sim \mathcal{N}(\mu_{style}, \Sigma_{style})$

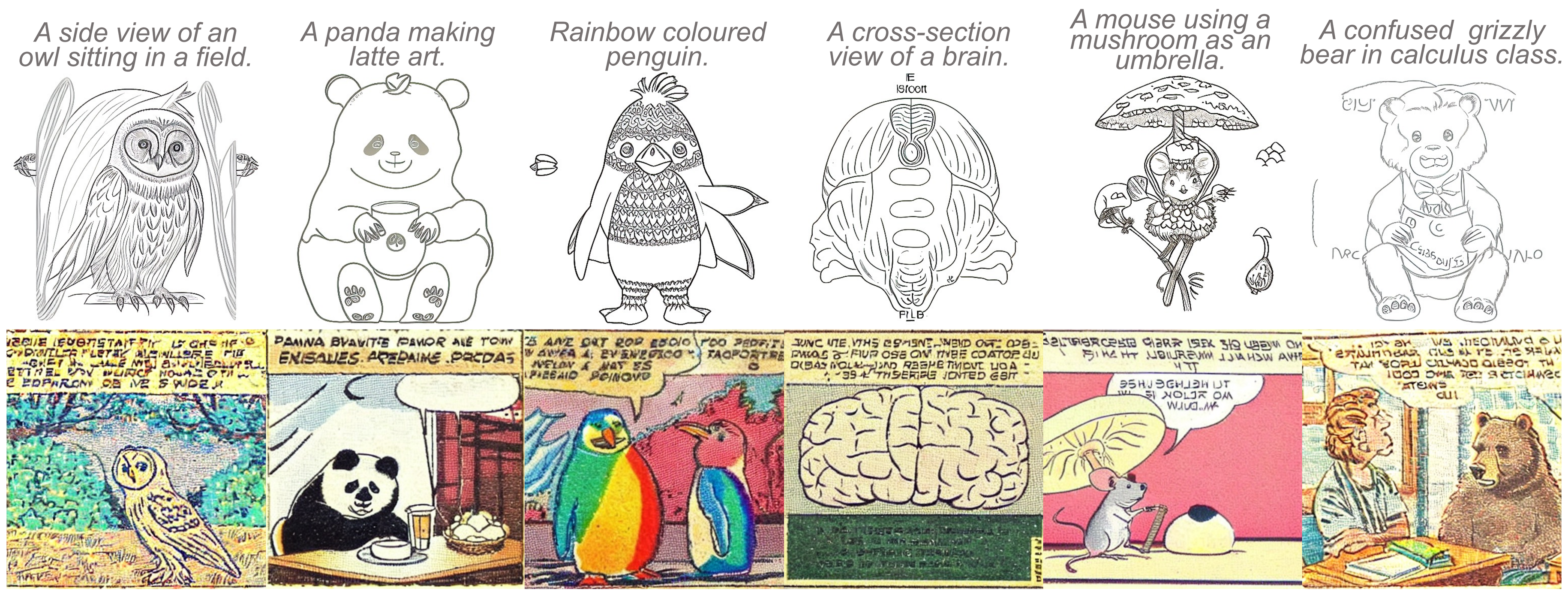
x_0 x_{333} x_{667} x_{1000}

We compute the style-specific noise parameters μ_{style} and Σ_{style} from **a small set of images of the desired style**. We use the finetuned model to denoise the initial noise $\hat{x}_{1000} \sim \mathcal{N}(\mu_{style}, \Sigma_{style})$.

We use our approach to fine-tune SD 1.5 [1] to different styles, *e.g.* anime sketches, or comics images.

Style 1, anime sketches [3]

Style 2, comics images [4]



$\hat{x}_{1000}$ $\hat{x}_{667}$ $\hat{x}_{333}$ $\hat{x}_0$ $\hat{x}_{1000}$ $\hat{x}_{667}$ $\hat{x}_{333}$ $\hat{x}_0$

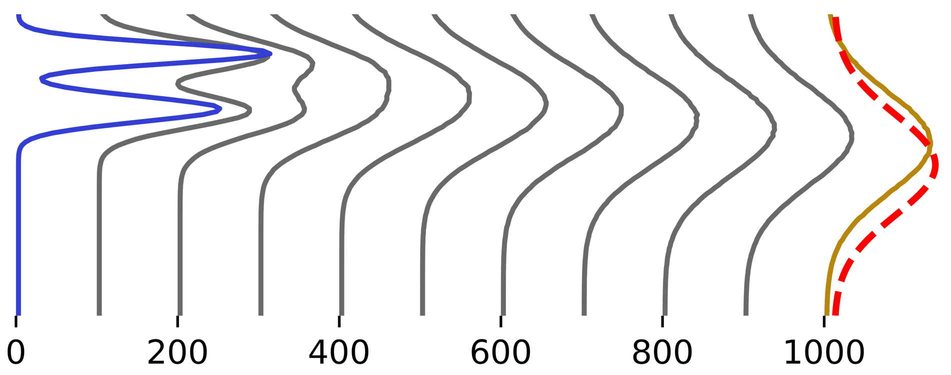
Exploiting the Signal-Leak Bias in Diffusion Models

Everaert M.N., Fitsios A., Bocchio M., Arpa S., Süsstrunk S., Achanta R. — JAN 4-8 WACV 2024 WAIKOLOA HAWAII

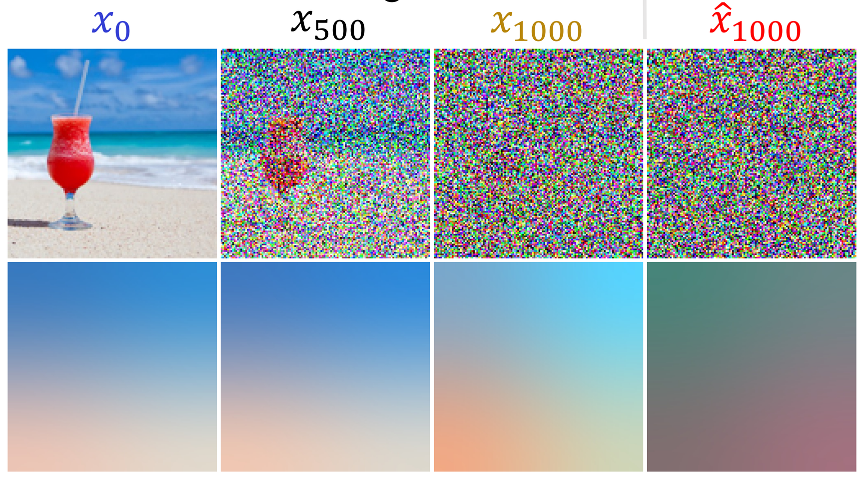
Diffusion models never fully corrupt images during training [5,6]:

$$x_{1000} = \sqrt{\bar{\alpha}_{1000}}\, x_0 + \sqrt{1-\bar{\alpha}_{1000}}\, \varepsilon \quad \text{with} \quad x_0 \sim p(x_0) \quad \text{and} \quad \varepsilon \sim \mathcal{N}(0_d, \mathrm{I}_{d\times d})$$
$$\approx 0.068\, x_0 \;\; + 0.998\, \varepsilon$$

However, the process of **generating images starts with pure noise** $\hat{x}_{1000} \sim \mathcal{N}(0_d, \mathrm{I}_{d\times d})$, oblivious of the **signal leak** $\sqrt{\bar{\alpha}_{1000}}\, x_0$ present in x_{1000} during training, **creating a bias**.



The diffusion model uses the signal-leak $\sqrt{\bar{\alpha}_{1000}}\, x_0$ in x_{1000} to deduce the **low-frequency information** about x_0. Using $\hat{x}_{1000} \sim \mathcal{N}(0_d, \mathrm{I}_{d\times d})$ **biases** the low-frequency components towards **medium values**.



Instead of retraining or finetuning [5,6,A] to remove this bias, we exploit it to our advantage by **including a signal-leak** $\sqrt{\bar{\alpha}_{1000}}\, \tilde{x}$ in $\hat{x}_{1000}$ **at inference time**, starting generating images from:

$$\hat{x}_{1000} = \sqrt{\bar{\alpha}_{1000}}\, \tilde{x} + \sqrt{1-\bar{\alpha}_{1000}}\, \varepsilon \quad \text{with} \; \tilde{x} \sim q(\tilde{x}) \; \text{and} \; \varepsilon \sim \mathcal{N}(0_d, \mathrm{I}_{d\times d})$$

With $q(\tilde{x}) = \mathcal{N}(\mu_{style}, \Sigma_{style})$, we exploit the bias to generate images $\hat{x}_0$ in the style we want:

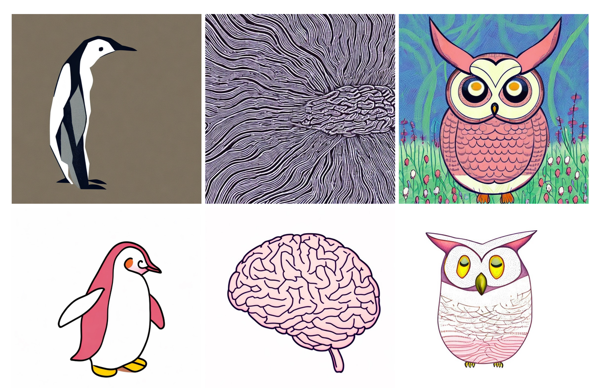
line-art model [7]

line-art model [7] **with ours**

SD 2.1 [1]

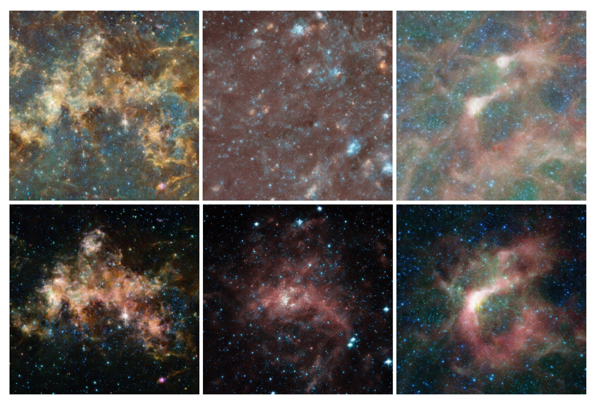
[...] in the style of line art, pastel colors, white background.

SD 2.1 [1] **with ours**



nasa space model [8]

nasa space model [8] **with ours**

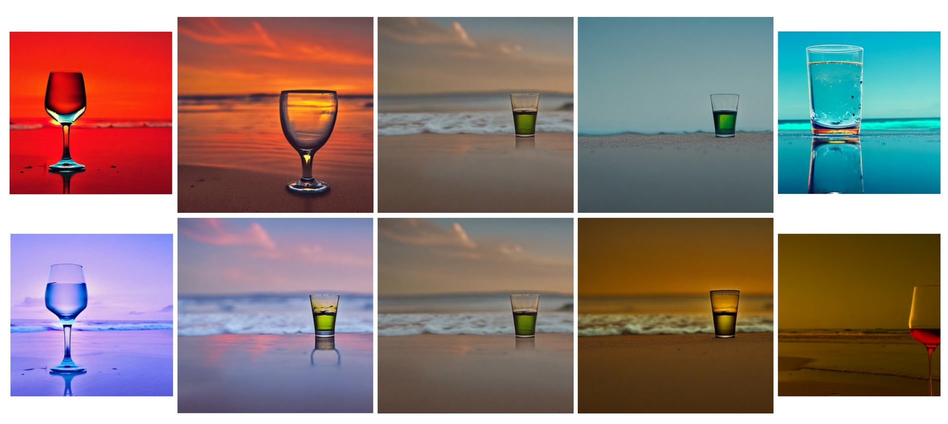


SD 2.1 [1]

A blue city at night, long exposure, orange and blue.

SD 2.1 [1] **with ours**

At inference time, we can control the low-frequency components of the generated images $\hat{x}_0$ by setting the desired ones (here, the mean color) in $\tilde{x}$:



References

[A] Everaert M.N. et al. "Diffusion in style." ICCV 2023.
[B] Everaert M.N. et al. "Exploiting the signal-leak bias in diffusion models." WACV 2024.
[1] Rombach R. et al. "High-resolution image synthesis with latent diffusion models." CVPR 2022.
[2] Nichol A. and Dhariwal P. "Improved denoising diffusion probabilistic models." ICML 2021.
[3] Taebum K. "Anime Sketch Colorization dataset." Kaggle dataset. 2018.
[4] *Simon* and *Kirby*. "48 Famous Americans." 1947.
[5] Guttenberg N. "Diffusion with Offset Noise." 2023
[6] Lin S. et al. "Common Diffusion Noise Schedules and Sample Steps are Flawed." WACV 2024.
[7] Karan D. "line-art" model. via huggingface.co/sd-concepts-library. 2022.
[8] *MatAlart*. "nasa space" model. Via huggingface.co/sd-dreambooth-library. 2022

ivrl.github.io/diffusion-in-style/

ivrl.github.io/signal-leak-bias/

Acknowledgement

These 2 works [A, B] were supported by Innosuisse grant 48552.1 IP-ICT.