Exploiting the Signal-Leak Bias in Diffusion Models

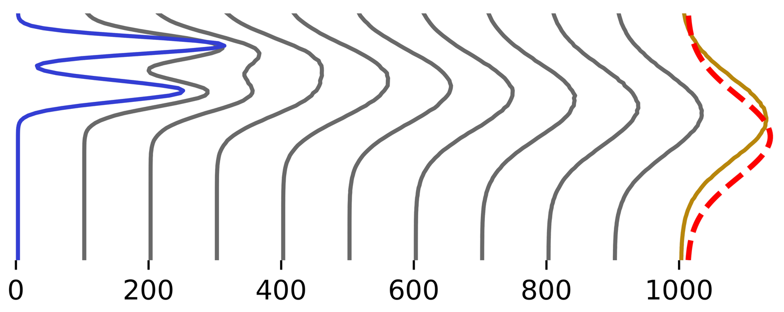
Martin Nicolas Everaert | Athanasios Fitsios | Marco Bocchio | Sami Arpa | Sabine Süsstrunk | Radhakrishna Achanta

Signal-leak bias

We can generate images in a **desired style** or with a more natural color distribution **without retraining** the diffusion model, by exploiting a **signal-leak bias** present in the model.

Common **diffusion models never fully corrupt images** during training [1,2]:
$x_T = \sqrt{\bar{\alpha}_T}\, x_0 + \sqrt{1-\bar{\alpha}_T}\, \varepsilon$ with $x_0 \sim p(x_0)$ and $\varepsilon \sim \mathcal{N}(0, I)$

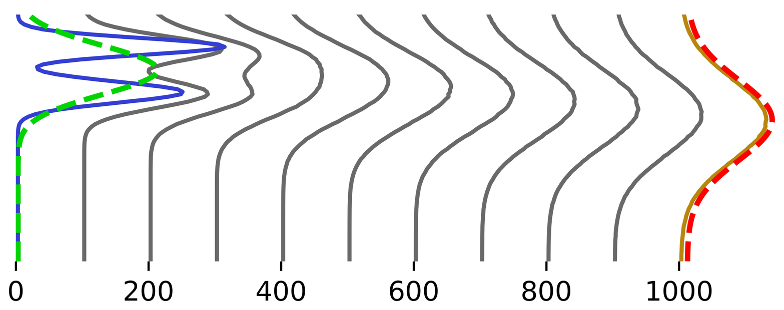
However, the process of **generating images starts with pure noise** $\hat{x}_T \sim \mathcal{N}(0, I)$, oblivious of the **signal leak** $\sqrt{\bar{\alpha}_T}\, x_0$ present in x_T during training, **creating a bias**.



$\hat{x}_T \sim \mathcal{N}(0, I)$
$\neq p(x_T)$

Instead of retraining or finetuning [1,2,3] to remove this bias, we exploit it to our advantage, generating images in the style we want.

We **include a signal-leak** $\sqrt{\bar{\alpha}_T}\, \tilde{x}$ in $\hat{x}_T$ **at inference time**, starting generating images from:
$\hat{x}_T = \sqrt{\bar{\alpha}_T}\, \tilde{x} + \sqrt{1-\bar{\alpha}_T}\, \varepsilon$ with $\tilde{x} \sim q(\tilde{x})$ and $\varepsilon \sim \mathcal{N}(0, I)$



$q(\tilde{x}) \approx p(x_0)$ realigns training and inference distributions.

Fixing style-adapted models

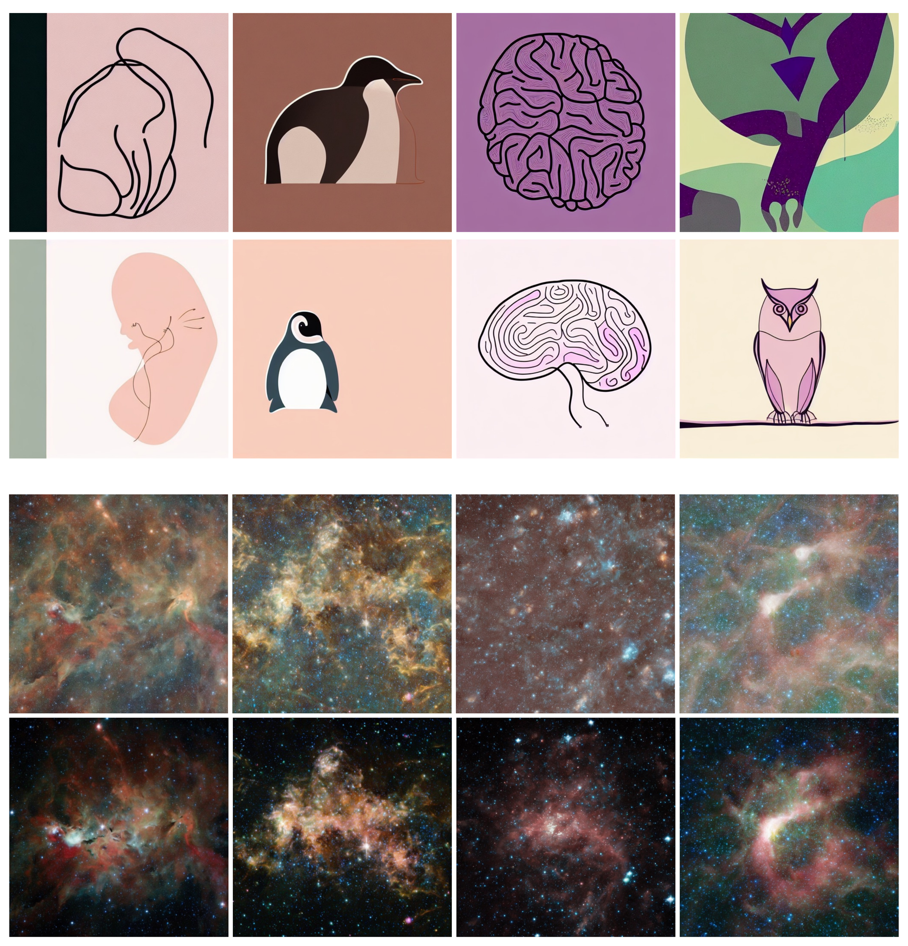
We obtain a distribution $q(\tilde{x})$ in the **pixel domain**, by approximating the distribution $p(x_0)$ as independent Gaussian distributions for each pixel.

"line-art" model [5]

"line-art" model [5] **with ours**

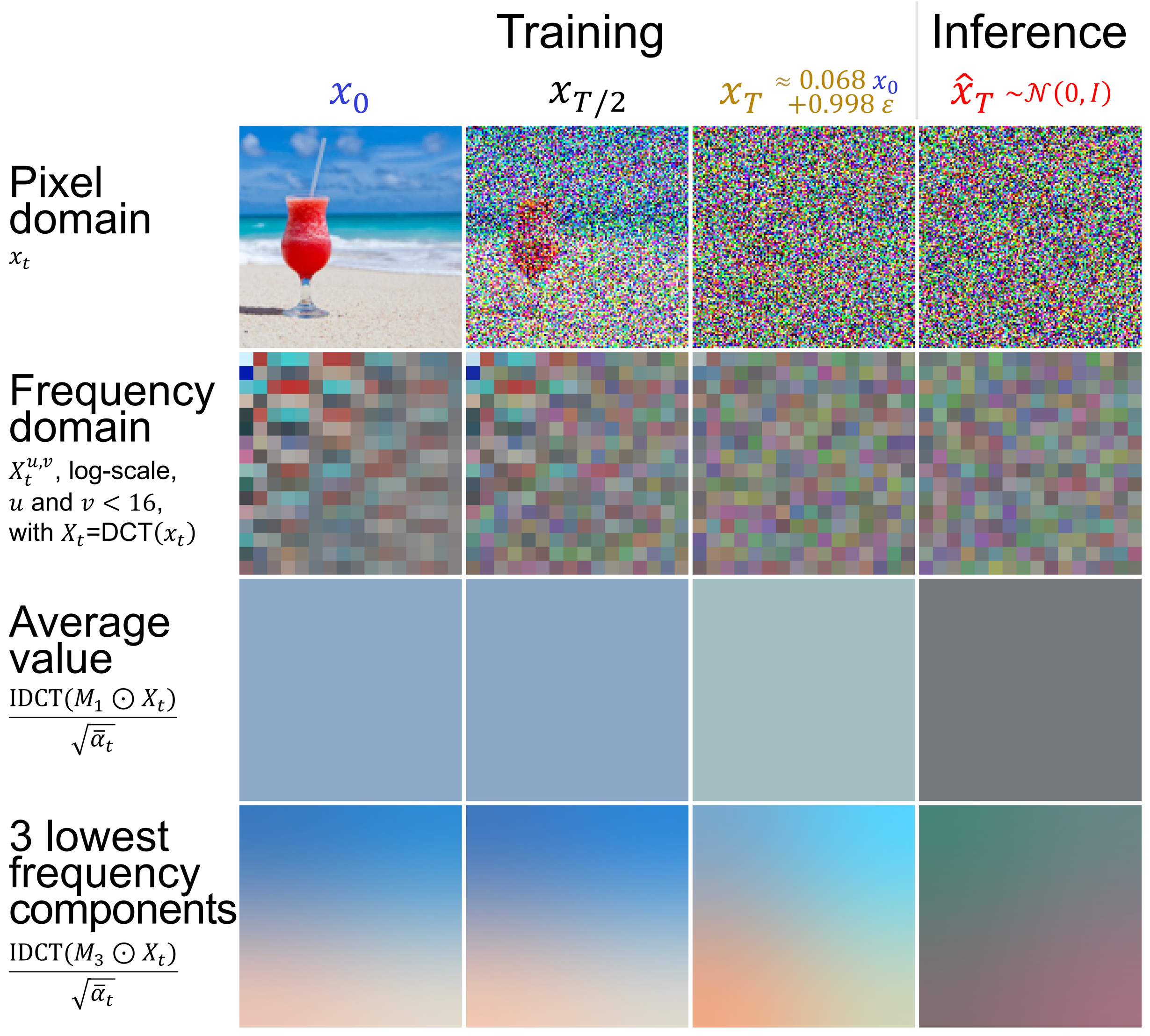
"nasa space" model [7]

"nasa space" model [7] **with ours**



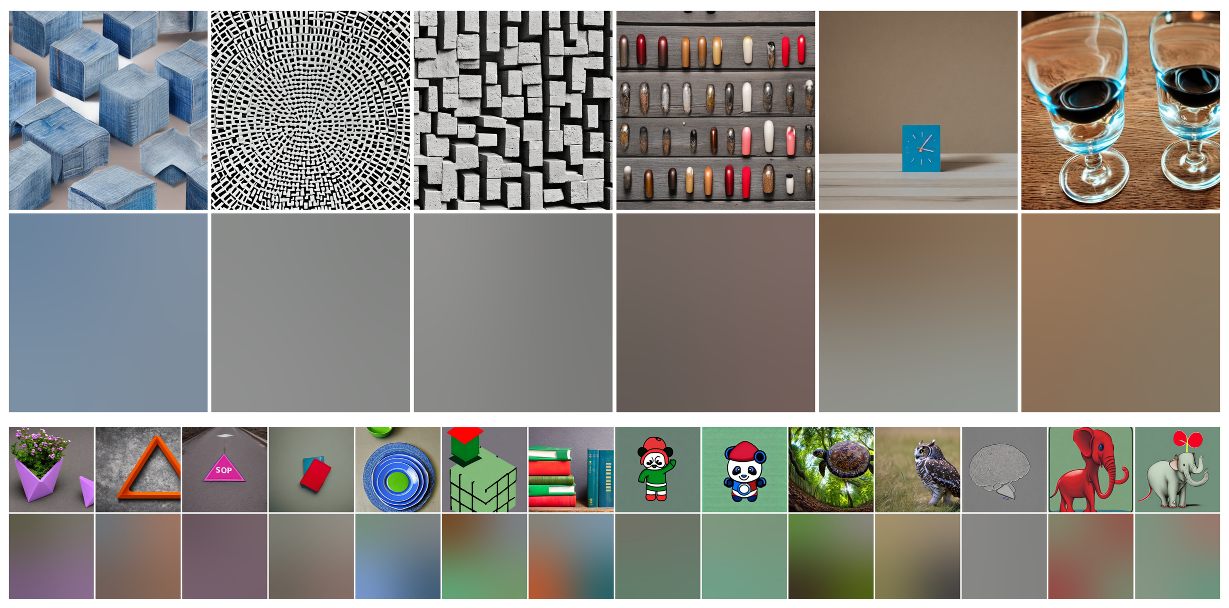
Better low-frequency components

The diffusion model uses the signal-leak $\sqrt{\bar{\alpha}_T}\, x_0$ to deduce the **low-frequency information** about x_0 from x_T. Using $\hat{x}_T \sim \mathcal{N}(0, I)$ **biases** the low-frequency components towards **medium values**.

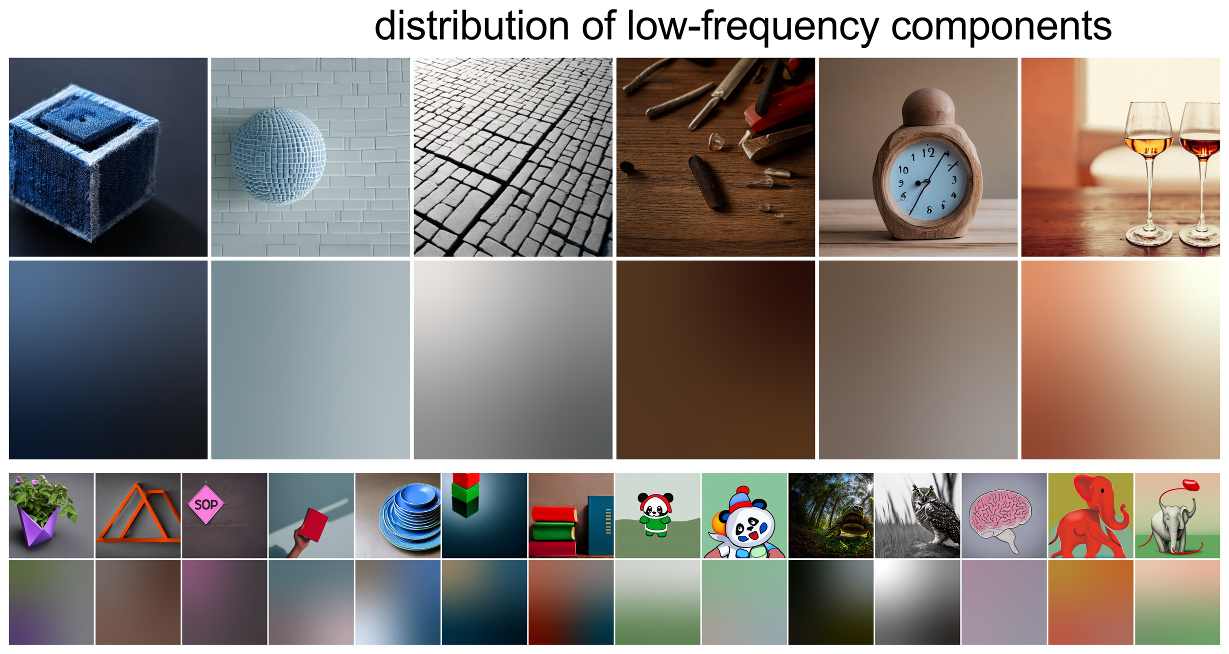


To avoid this, we additionnally **model the low-frequency components**, estimating their mean and covariance, and obtain a distribution $q(\tilde{x}) \approx p(x_0)$.

Original results SD 2.1 [4] → greyish images with low contrast or variation of colors



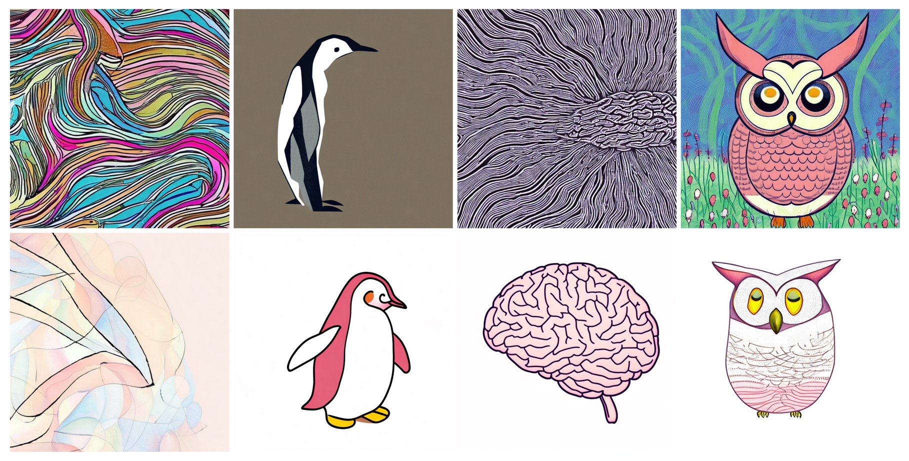
Our results SD 2.1 [4] **with ours** → more varied and natural distribution of low-frequency components



Style-adaptation with the original diffusion model

SD 2.1 [4]
"[…] in the style of line art, pastel colors, white background."

SD 2.1 [4] **with ours**

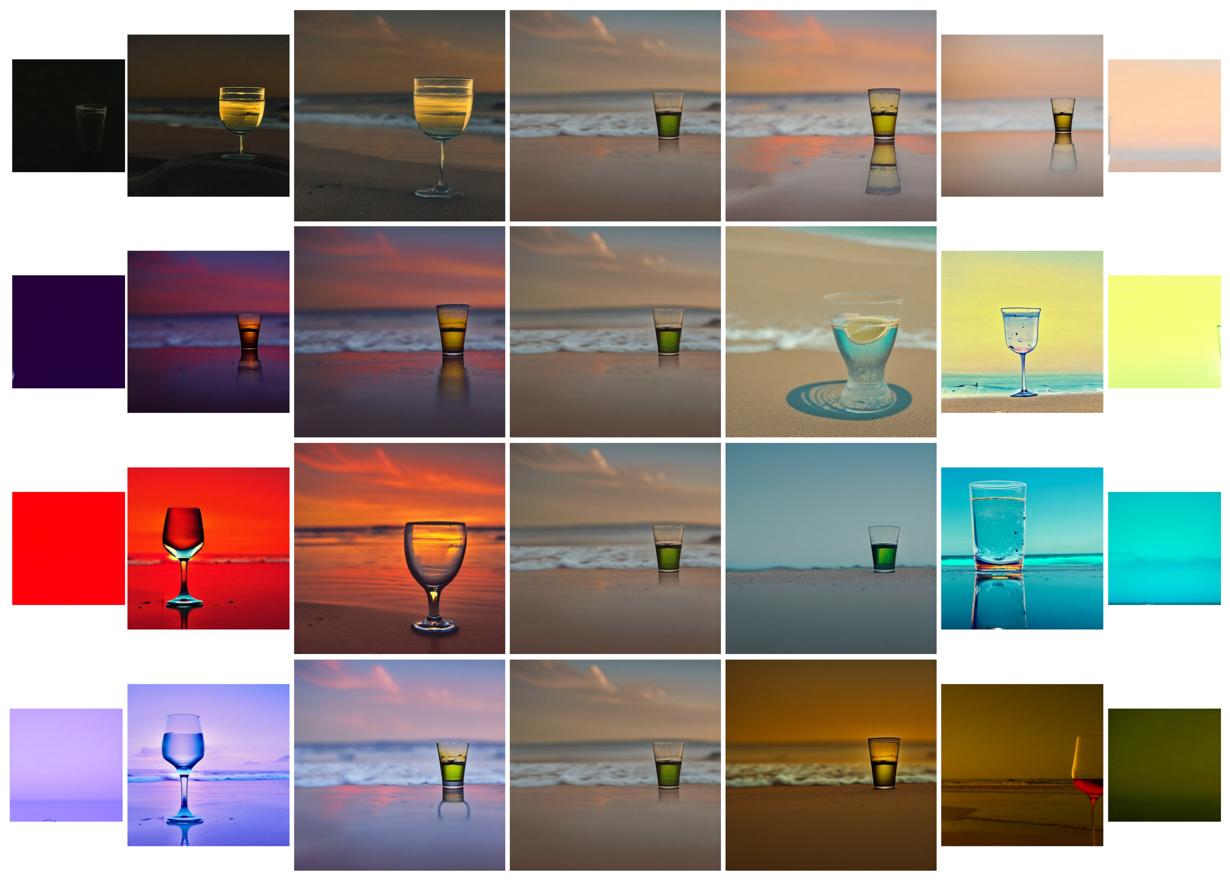


SD 2.1 [4]
"A blue city at night, long exposure, orange and blue."

SD 2.1 [4] **with ours**

More control on low-frequency components

Setting manually the signal-leak $\sqrt{\bar{\alpha}_T}\, \tilde{x}$ in $\hat{x}_T$ → control on the low-frequency components (*e.g.*, the mean color of the generated images)



References

"line-art" model [5]: Stable Diffusion v1.4 finetuned with Textual Inversion [5,6] on 7 line-art images [5] (bright background, pastel colors)
"nasa space" model [7]: Stable Diffusion v2 finetuned with DreamBooth [7,8] on 24 photos of astronomical phenomena [7]
Blue city at night: using 9 images from https://unsplash.com/collections/67793987 (Credits: Unsplash, @borkography)

[1] Guttenberg. Diffusion with Offset Noise. 2023
[2] Lin et al. Common Diffusion Noise Schedules and Sample Steps are Flawed. arXiv 2023
[3] Everaert et al. Diffusion in Style. ICCV 2023
[4] Stability AI. Stable Diffusion 2.1. 2022 + Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022
[5] Karan. "line-art" model. https://huggingface.co/sd-concepts-library/line-art. 2022
[6] Gal et al. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. ICLR 2023
[7] MatAIart. "nasa space" model. https://huggingface.co/sd-dreambooth-library/nasa-space-v2-768. 2022
[8] Ruiz et al. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. CVPR 2023

This work is supported by Innosuisse grant 48552.1 IP-ICT.

Project website:

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