Exploiting the Signal-Leak Bias in Diffusion Models (Supplementary material)

Martin Nicolas Everaert[1] Athanasios Fitsios[1,2] Marco Bocchio[2]
Sami Arpa[2] Sabine Süsstrunk[1] Radhakrishna Achanta[1]
[1]School of Computer and Communication Sciences, EPFL, Switzerland [2]Largo.ai, Lausanne, Switzerland
Project page: https://ivrl.github.io/signal-leak-bias/

This supplementary document reproduces some figures of the manuscript in higher resolution, shows additional figures, includes a discussion on failure cases, and contains supporting statistics further demonstrating that our approach allows us to generate images with more varied low-frequency components.

The reader is encouraged to look at this supplementary document in a digital format, in order to be able to zoom into the generated images.

Contents

1. Improving style for style-specific models

1.1. Style 1: Pokemon BLIP captions dataset

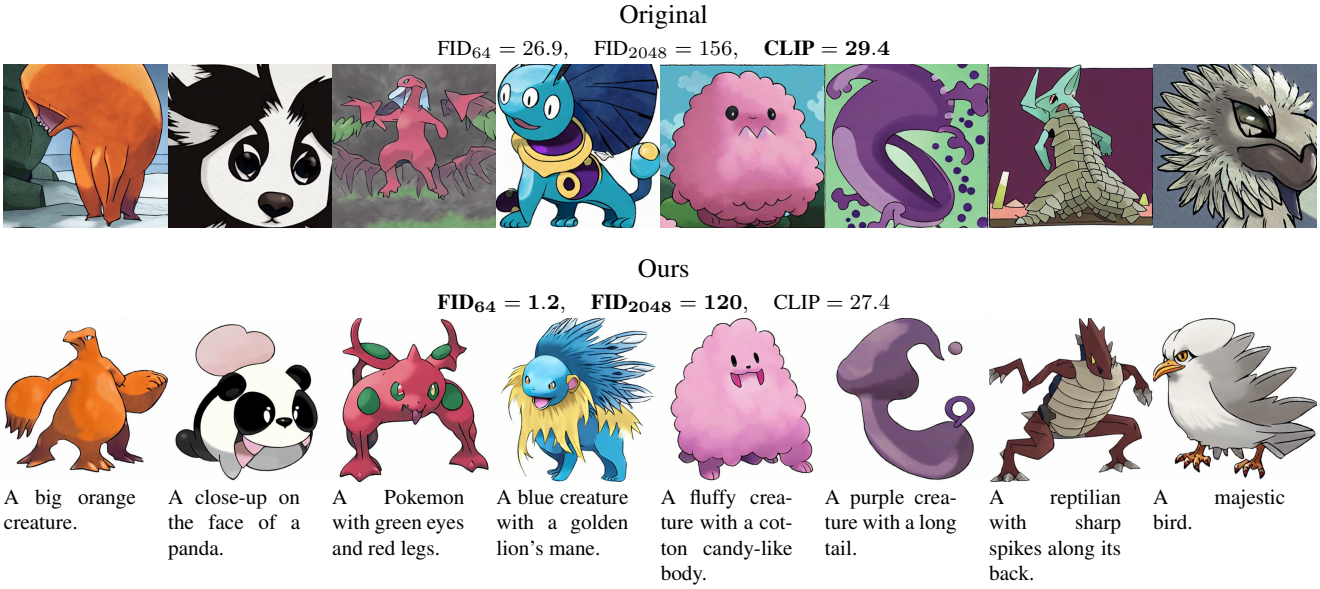


Figure 1. This figure replicates Figure 3a of the manuscript, in higher resolution and with 4 additional columns. All images are obtained with the Pokemon-LoRA model that was trained on the Pokemon BLIP captions dataset.
While the images from the Pokemon BLIP captions dataset all have a white background, this is not reproduced correctly when using white noise as initial latents (top row).
In contrast, using initial latents composed of white noise and a signal leak as we propose (bottom row) adheres to the desired style more faithfully, and this without additional training.
The textual prompts used to generate the images are provided at the bottom of each column.

1.2. Style 2: line-art images

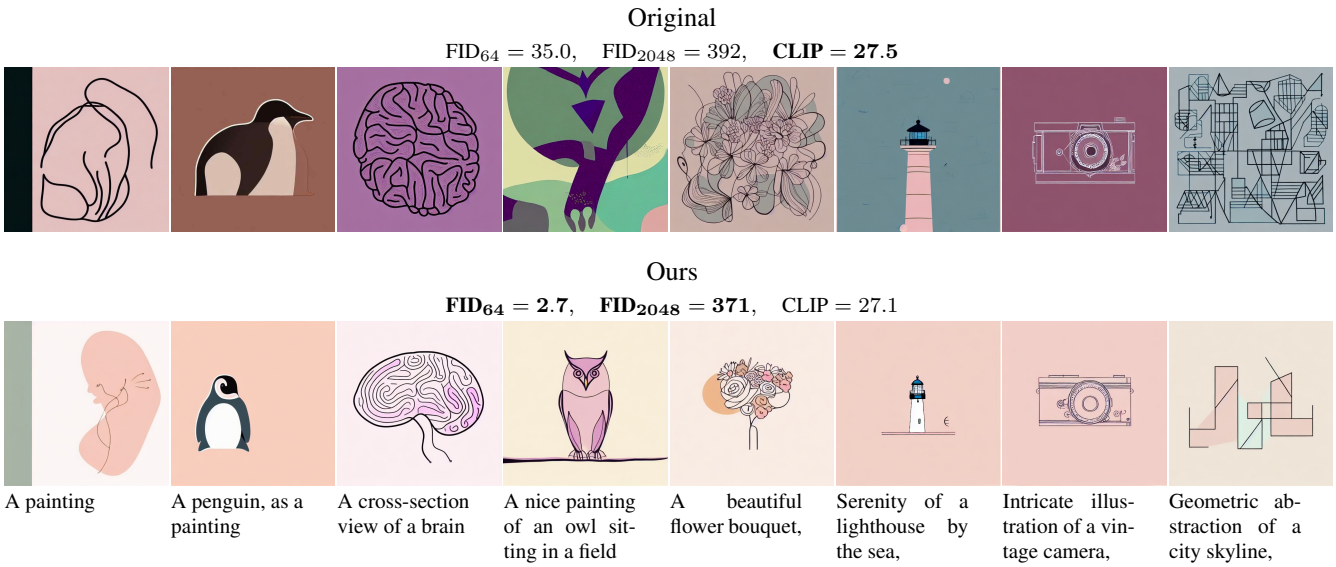
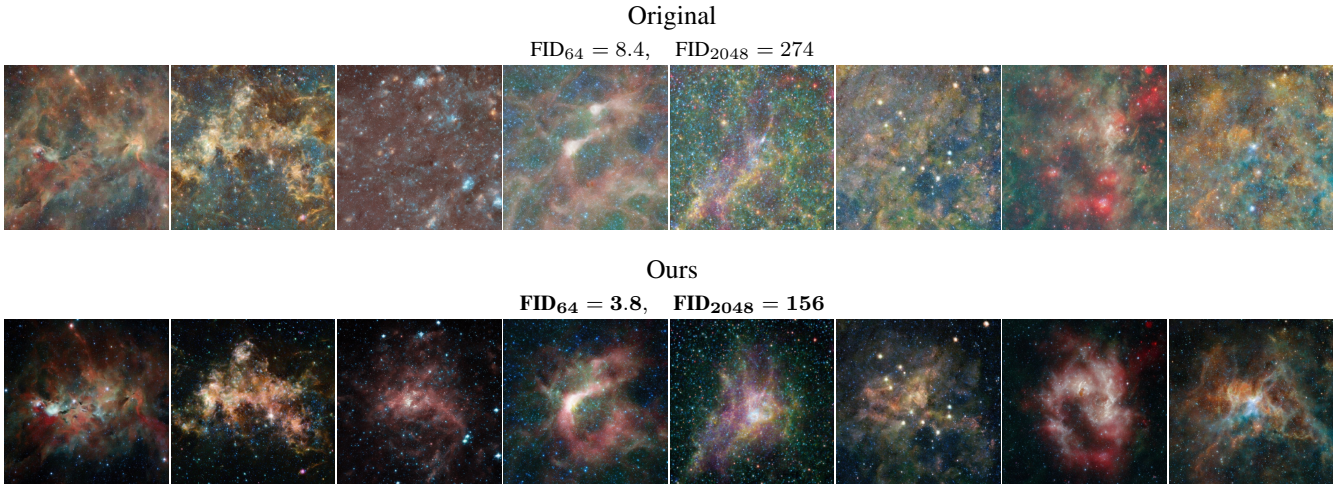


Figure 2. This figure replicates Figure 3b of the manuscript, in higher resolution and with 4 additional columns. All images are obtained with Stable Diffusion v1.4 using the <line-art> concept learned with Textual Inversion from 7 line-art images.
While the 7 line-art images all have a bright background and pastel colors, this is not reproduced correctly when using white noise as initial latents (top row).
In contrast, using initial latents composed of white noise and a signal leak as we propose (bottom row) adheres to the desired style more faithfully, and this without additional training.
The textual prompts used to generate the images are provided at the bottom of each column.

1.3. Style 3: images of astronomical phenomena



Textual prompt for all images: Nasa style

Figure 3. This figure replicates Figure 3c of the manuscript, in higher resolution and with 4 additional columns. All images are obtained from the prompt "Nasa style" with the NASA-space-V2 model that was trained on the 24 images of astronomical phenomena.
While the 24 training images are all dark, this is not reproduced correctly when using white noise as initial latents (top row).
In contrast, using initial latents composed of white noise and a signal leak as we propose (bottom row) adheres to the desired style more faithfully, and this without additional training.

2. Improving style for non-style-specific models

2.1. Style 1: Pokemon BLIP captions dataset

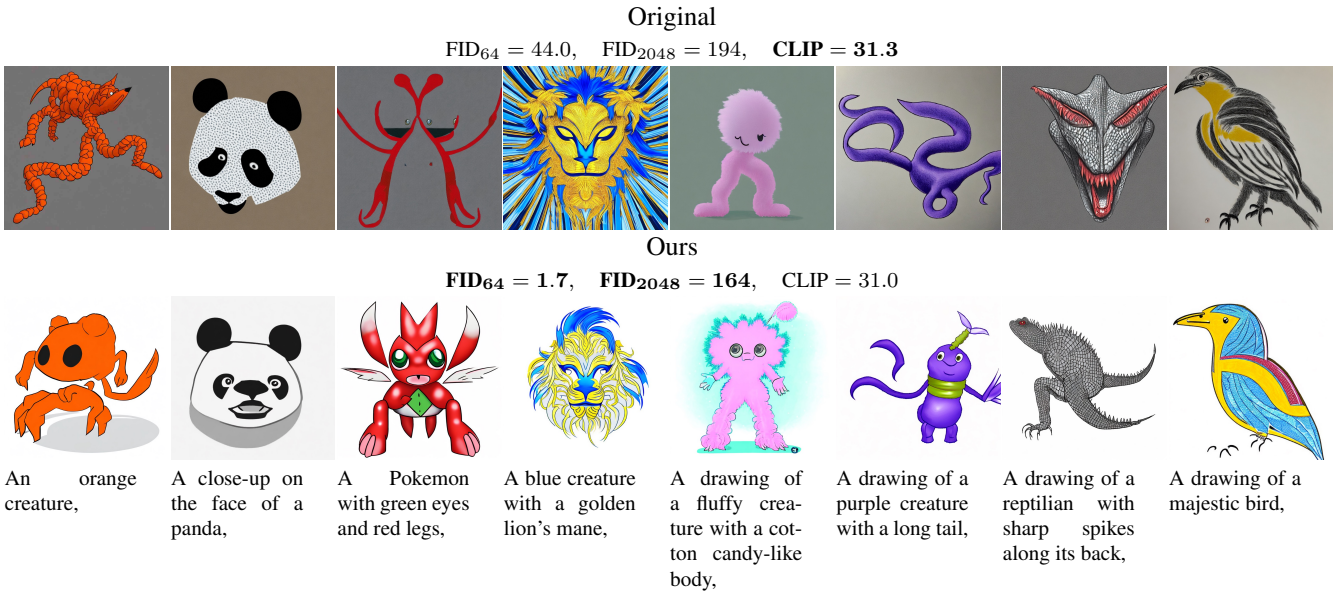


Figure 4. This figure replicates Figure 4a of the manuscript, in higher resolution and with 4 additional columns. All images are obtained with the original Stable Diffusion V2.1 model, which has not been fine-tuned for a specific style.
Here, we aim to mimic the style of images from Pokemon BLIP captions dataset, and hence add "in the style of Satoshi Tajiri, white background." inside each prompt.
While the images from the Pokemon BLIP captions dataset all have a white background, this is not reproduced correctly when using white noise as initial latents (top row).
In contrast, using initial latents composed of white noise and a signal leak as we propose (bottom row) adheres to the desired style more faithfully, and this without additional training.
The textual prompts used to generate the images are provided at the bottom of each column.

2.2. Style 2: line-art images

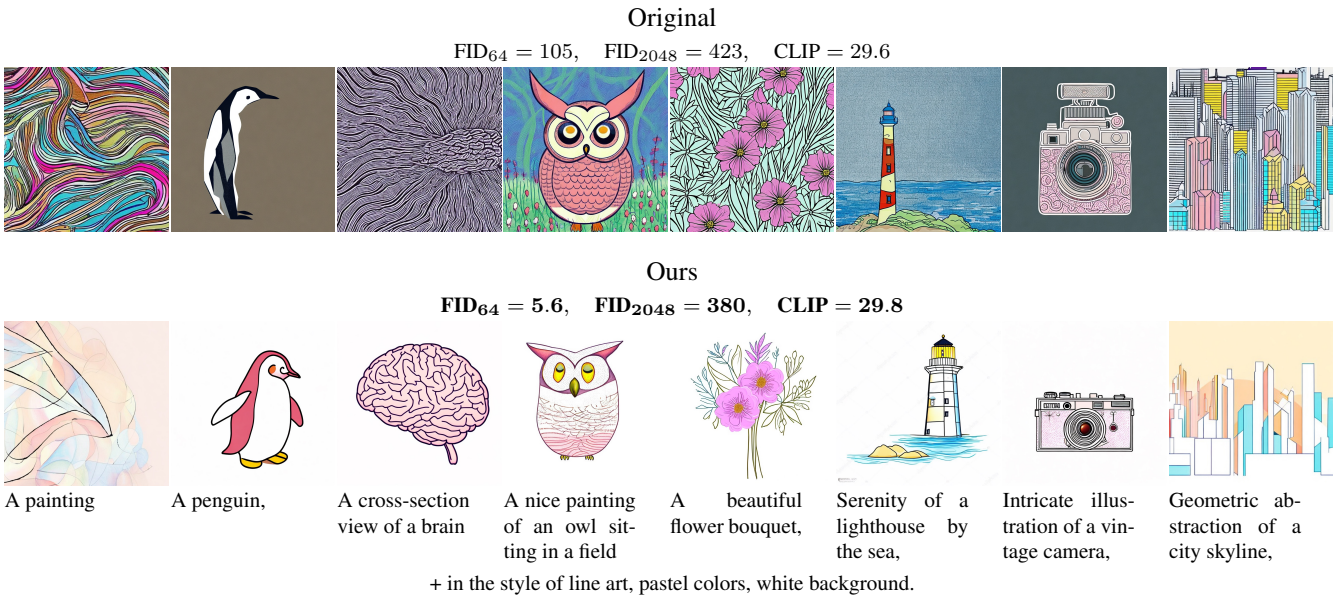


Figure 5. This figure replicates Figure 4b of the manuscript, in higher resolution and with 4 additional columns. All images are obtained with the original Stable Diffusion V2.1 model, which has not been fine-tuned for a specific style.
Here, we aim to mimic the style of the 7 line-art images, and hence add "in the style of line art, pastel colors, white background." inside each prompt.
While the 7 line-art images all have a bright background and pastel colors, this is not reproduced correctly when using white noise as initial latents (top row).
In contrast, using initial latents composed of white noise and a signal leak as we propose (bottom row) adheres to the desired style more faithfully, and this without additional training.
The textual prompts used to generate the images are provided at the bottom of each column.

2.3. Style 3: images of astronomical phenomena

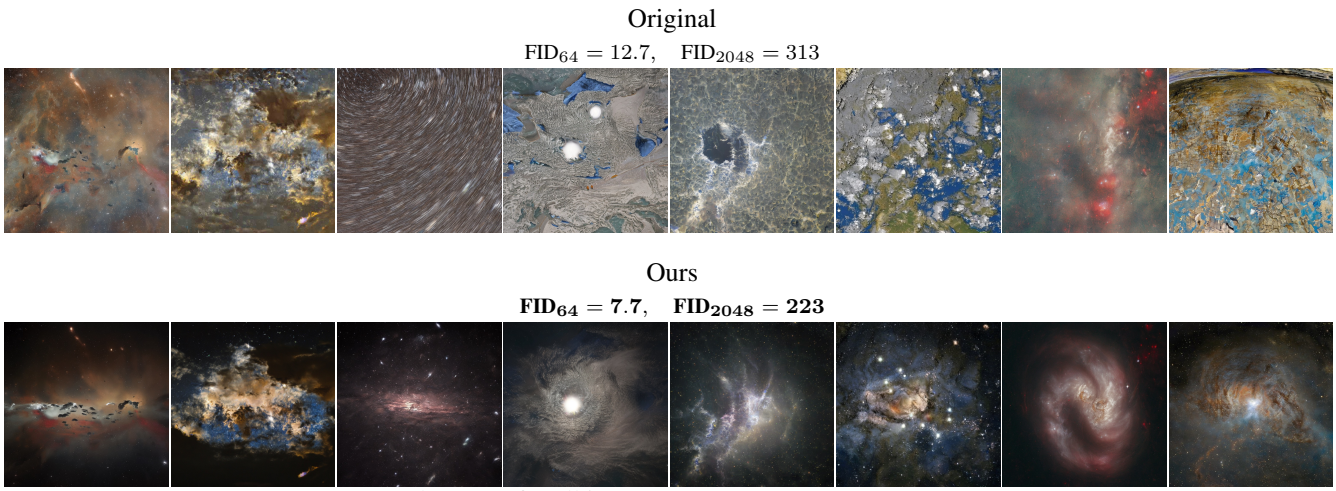


Figure 6. This figure replicates Figure 4c of the manuscript, in higher resolution and with 4 additional columns. All images are obtained with the original Stable Diffusion V2.1 model, which has not been fine-tuned for a specific style.
Here, we aim to mimic the style of the 24 images of astronomical phenomena, and hence generate images from the textual prompt "A photo of the sky made by the Nasa." inside each prompt.
While the 24 training images are all dark, this is not reproduced correctly when using white noise as initial latents (top row).
In contrast, using initial latents composed of white noise and a signal leak as we propose (bottom row) adheres to the desired style more faithfully, and this without additional training.

2.4. Styles 4 and 5: horizontal wave pattern (ablation/failure)

In this Section, we experiment with additional styles, corresponding to a horizontal wave pattern with 48 periods, as in Figures 7 and 8. The goal here is to study cases where our approach to "improving style for non-style-specific model" would fail:

- Case 1: As explained in the manuscript, noise mostly contaminates the high-frequency components, but only a little the low-frequency components. Consequently, the signal leak $\sqrt{\bar{\alpha}_T}\tilde{\boldsymbol{x}}$ that we include in the initial latent will mostly affect the low-frequency components of the generated images. As a consequence, we can foresee that if the characteristics of a style are related to high-frequency components only, our approach might not reproduce the style faithfully.
- Case 2: Our pixel-domain model, which we describe in Sections 4.2.1, 5.1, and 5.2 in the manuscript, computes statistics pixel-wise. This model might not be adapted to every style. For instance, if the style is such that pixel-wise

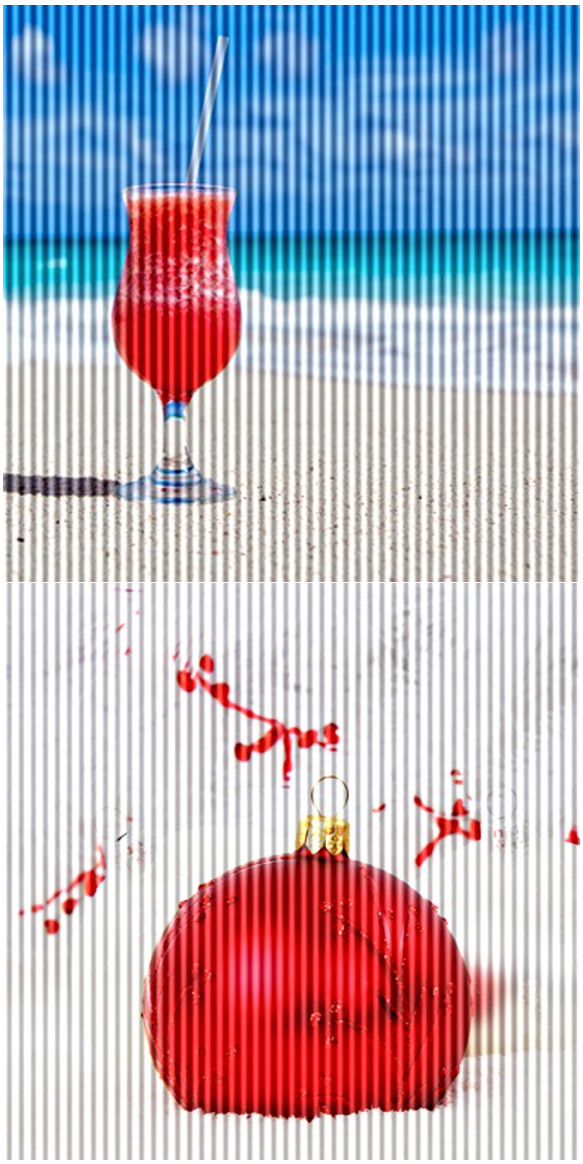


Figure 7. Style 4: horizontal wave pattern is aligned across images.

Figure 8. Style 5: horizontal wave pattern is not aligned across images

Credit for original images: Pixabay, Pixabay

statistics $\boldsymbol{\mu}, \boldsymbol{\sigma}$ of a set of stylized images are the same as the ones of a set of non-stylized images, then the signal leak $\sqrt{\bar{\alpha}_T}\tilde{\boldsymbol{x}}$ with our pixel-domain model would not contain any meaningful information to bias image generation toward the style.

To study how our approach performs on these 2 cases, we generate two new styles with a high-frequency horizontal wave pattern as in Figures 7 and 8. For style 4, we generated images such that the 48 periods are aligned within our set of target images. Therefore, statistics $\boldsymbol{\mu}, \boldsymbol{\sigma}$ computed with our pixel-domain model "see" the pattern of the style. For style 5, we generated images such that the 48 periods do not align across images of our set. In particular, we can think intuitively that dark and light vertical lines compensate for each other, causing the computed statistic $\boldsymbol{\mu}$ not to contain any information about the style.

We applied our method with these two new styles, and show the qualitative results in Figure 9. As we can see, our method fails to reproduce the styles correctly. For style 4, while our method managed to influence the generated images toward the correct style, the generated horizontal wave pattern is not regular and not consistent across all generated images. For style 5, it appears visually that our approach fails to influence the image generation towards the desired style. These observations are coherent with the intuition described in cases 1 and 2.

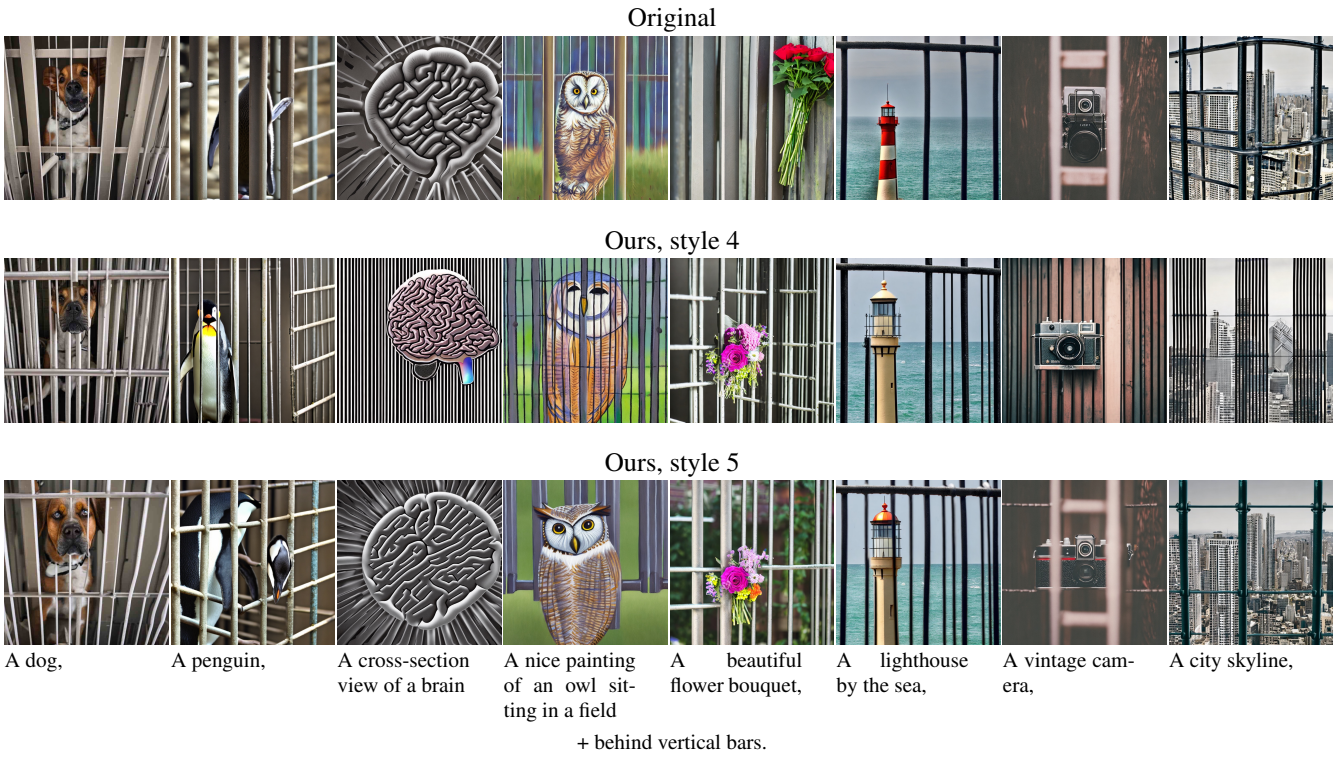


Figure 9. All images are obtained with the original Stable Diffusion V2.1 model, which has not been fine-tuned for a specific style.
Here, we aim to mimic the "horizontal wave pattern" described above (styles 4 and 5).
The specific characteristic of the style, *i.e.* the 48 horizontal cosine wave pattern is not reproduced correctly, both when using white noise as initial latents (top row), or using initial latents composed of white noise and our signal leak (second and third rows).

3. Generating images with more varied low-frequency components

3.1. Additional images

The images in Figure 5 of the manuscript are generated from the 1^{st}, 26^{th}, 51^{st}, ..., 151^{st}, 176^{th} prompts of the DrawBench benchmark, *i.e*.

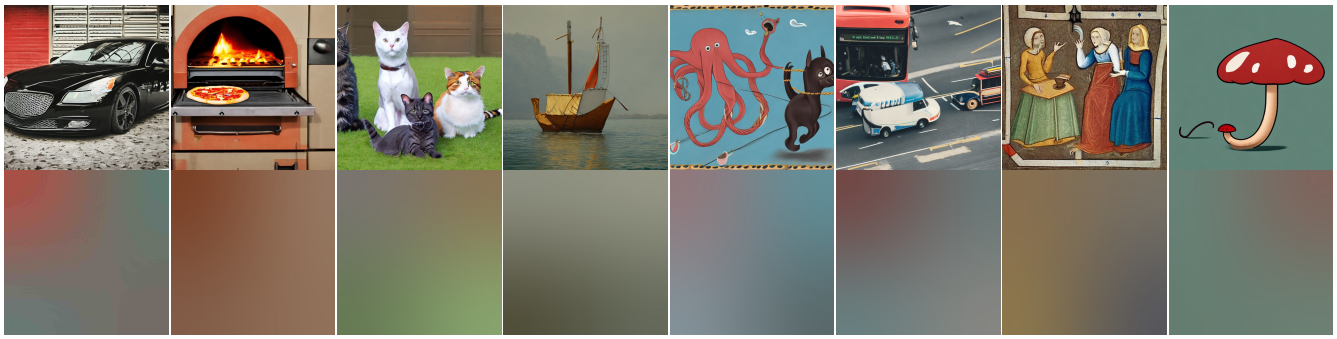
- A red colored car.
- A horse riding an astronaut.
- Two cats and three dogs sitting on the grass.
- A large motor vehicle carrying passengers by road, typically one serving the public on a fixed route and for a fare.
- A pear cut into seven pieces arranged in a ring.
- A couch on the left of a chair.
- A yellow and black bus cruising through the rainforest.
- A 1960s poster warning against climate change.

Additional images are shown below in Figure 10. These images are generated from the 2^{nd}, 27^{th}, 52^{nd}, ..., 152^{nd}, 177^{th} prompts of the DrawBench benchmark, *i.e*.

- A black colored car.
- A pizza cooking an oven.
- Three cats and one dog sitting on the grass.
- A small vessel propelled on water by oars, sails, or an engine.
- A donkey and an octopus are playing a game. The donkey is holding a rope on one end, the octopus is holding onto the other. The donkey holds the rope in its mouth. A cat is jumping over the rope.
- A car on the left of a bus.
- A medieval painting of the wifi not working.
- Illustration of a mouse using a mushroom as an umbrella.

Original

$\text{FID}_{64} = 2.65, \quad \text{FID}_{2048} = 192, \quad \mathbf{CLIP = 32.2}$



Ours

$\mathbf{FID_{64} = 2.64}, \quad \mathbf{FID_{2048} = 187}, \quad \text{CLIP} = 31.8$

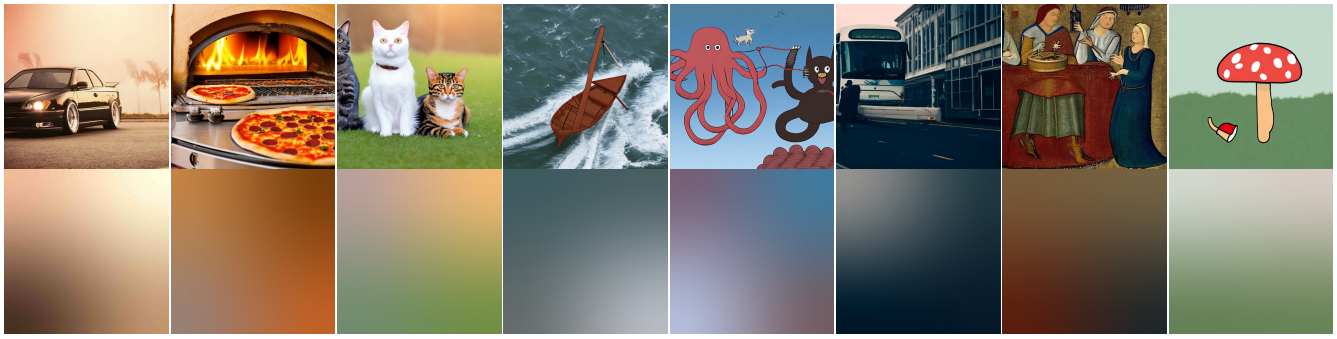


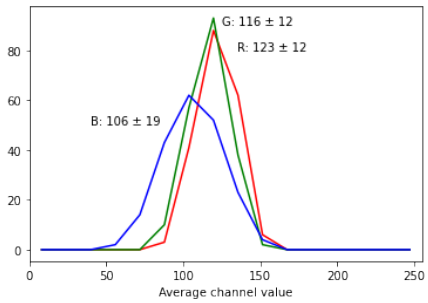
Figure 10. This figure mimics Figure 5 of the manuscript, with 8 different prompts. All 16 images are generated with the original Stable Diffusion V2.1 model. Below each generated image, we show a low-pass filtered version of it.
Using white noise as initial latents (top) biases the generated images to have medium low-frequency components, *i.e*. images tend to be greyish, with medium brightness and little variation of colors inside each image.
In contrast, using initial latents composed of white noise and a signal leak with more varied low-frequency components as we propose (bottom) generates images with more varied low-frequency components, and this without additional training. Images in the bottom have more varied colors and brightness across images and inside each image.

3.2. Image statistics

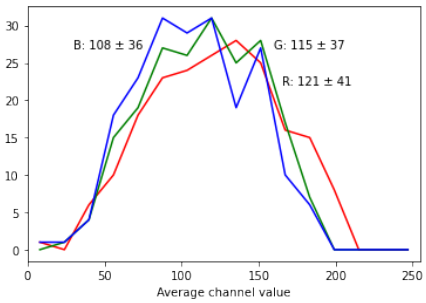
In Section 5.3 and Figure 5 of the manuscript, as well as in Figure 10 here, we present evidence that incorporating a signal leak in the initial latents generates images with more diverse low-frequency components. To further support this claim, we conducted an analysis of several basic image statistics, including average pixel values (Figure 11), average luminance (Figure 12), and contrast (Figure 13), for four sets of images: 200 images generated from white noise, 200 images generated from our approach, the first 1000 images from the COCO 2014 validation set, and 323 images from the LAION-6+ dataset. For each statistic, we provide the mean and standard deviation represented as mean $\pm$ std across a histogram.

These figures further confirm that initiating the denoising process from white noise tends to bias the generated images towards medium low-frequency components. These images exhibit characteristics such as a greyish appearance, medium brightness, and limited color variation within each image.

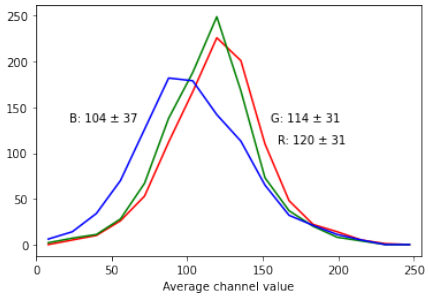
In contrast, the results obtained using our approach have more diverse low-frequency components. The histograms are wider, and the standard deviations are higher. Consequently, the generated images exhibit a more natural distribution of low-frequency components, including a greater variety of average colors, luminance levels, and contrast.



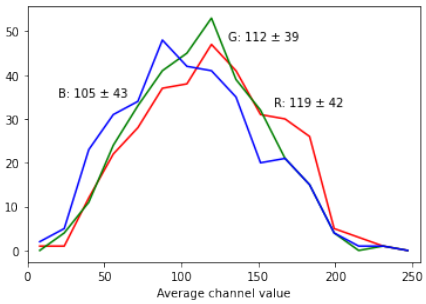
(a) Original SD 2.1 results



(b) SD 2.1 results with our apporach

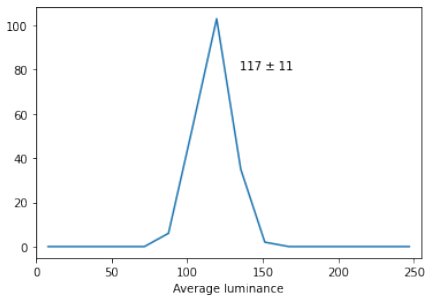


(c) 1000 first images of COCO Validation 2014

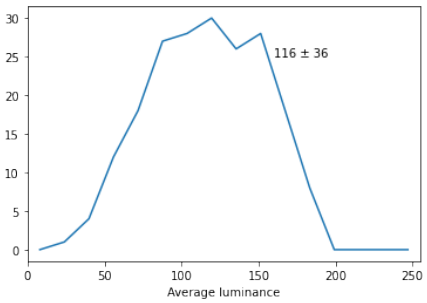


(d) 323 images from LAION-6+

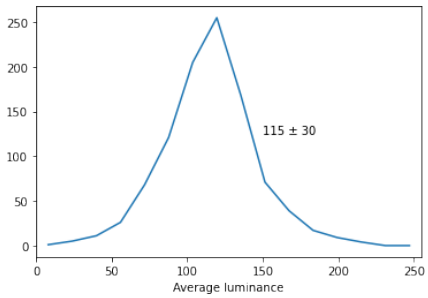
Figure 11. Average value of each channel (R, G, B).
We compute the average value of each channel (R, G, B) for each image in the 4 different sets of images. The histograms represent the distribution of mean RGB values computed for each image in the respective sets. The four subfigures in the figure compare the average pixel values of 200 images obtained from Stable Diffusion 2.1 with white noise (a), the average pixel values of 200 images obtained from Stable Diffusion 2.1 with our approach (b), the average pixel values of the first 1000 images from the COCO Validation 2014 dataset (c), and the average pixel values of 323 images from the LAION-6+ dataset (d).
Images generated from white noise using Stable Diffusion (a) cover only a moderate range of colors.
In contrast, images generated using our approach (b) cover a wider range of colors, similar to real images (c and d).



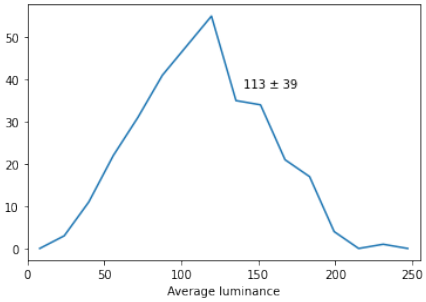
(a) Original SD 2.1 results



(b) SD 2.1 results with our apporach

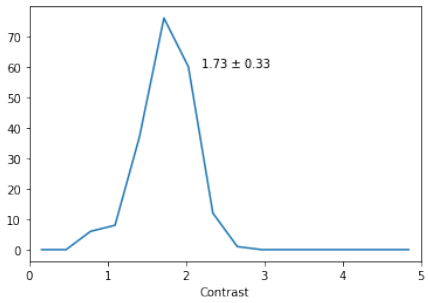


(c) 1000 first images of COCO Validation 2014

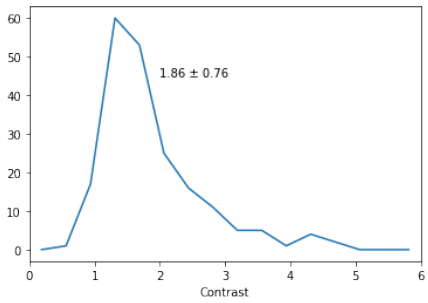


(d) 323 images from LAION-6+

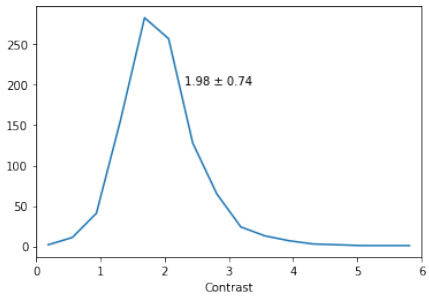
Figure 12. Average luminance.
We compute the average luminance for each image in the 4 different sets of images. The histograms represent the distribution of the average luminance computed for each image in the respective sets. The four subfigures in the figure compare the average luminance of 200 images obtained from Stable Diffusion 2.1 with white noise (a), the average luminance of 200 images obtained from Stable Diffusion 2.1 with our approach (b), the average luminance of the first 1000 images from the COCO Validation 2014 dataset (c), and the average luminance of 323 images from the LAION-6+ dataset (d).
Images generated from white noise using Stable Diffusion (a) cover only a moderate range of luminance.
In contrast, images generated using our approach (b) cover a wider range of luminance, similar to real images (c and d).



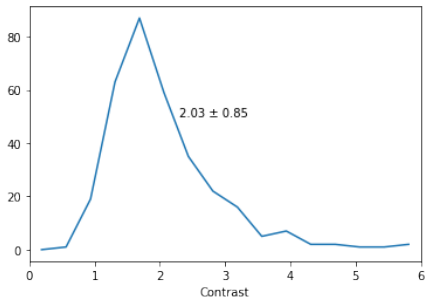
(a) Original SD 2.1 results



(b) SD 2.1 results with our apporach



(c) 1000 first images of COCO Validation 2014



(d) 323 images from LAION-6+

Figure 13. Contrast.
Here, the contrast of an image is defined as the difference between the 2nd and 98th percentiles of the luminance, divided by the average luminance.
We compute the contrast for each image in the 4 different sets of images. The histograms represent the distribution of the contrast computed for each image in the respective sets. The four subfigures in the figure compare the contrast in 200 images obtained from Stable Diffusion 2.1 with white noise (a), the contrast in 200 images obtained from Stable Diffusion 2.1 with our approach (b), the contrast of the first 1000 images from the COCO Validation 2014 dataset (c), and the contrast of 323 images from the LAION-6+ dataset (d).
Images generated from white noise using Stable Diffusion (a) cover only a moderate range of contrast.
In contrast, images generated using our approach (b) cover a wider range of contrast, similar to real images (c and d).