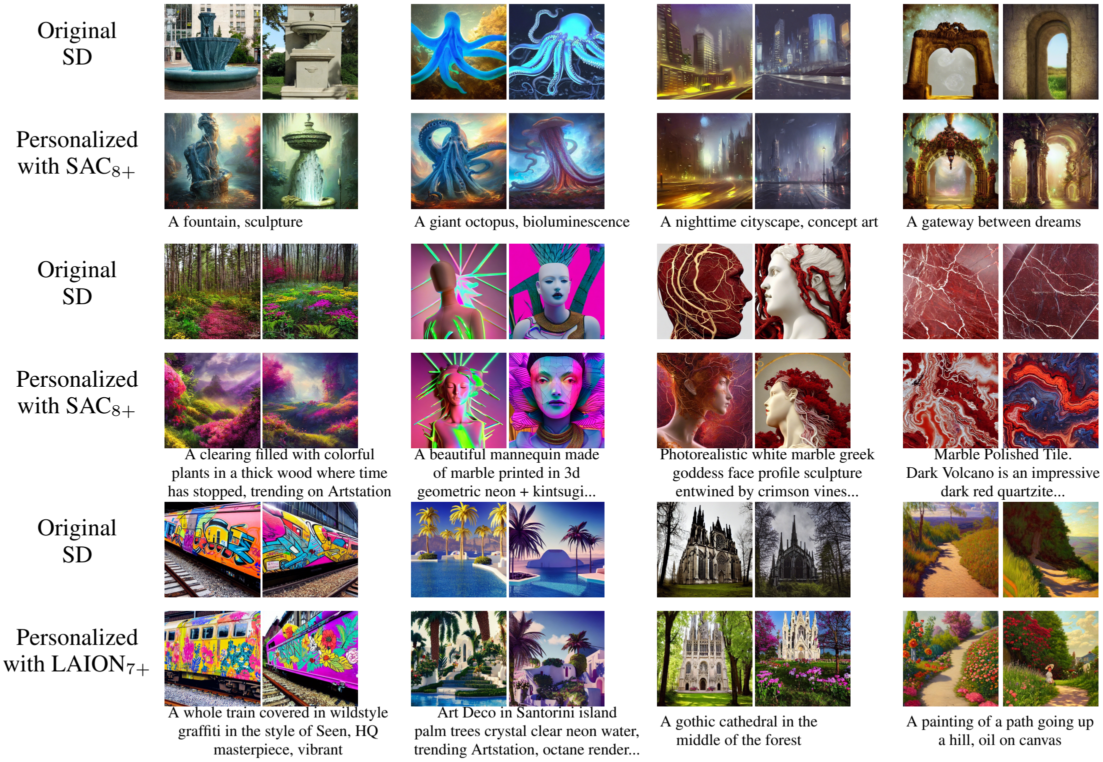
Personalizing Text-to-Image Generation via Aesthetic Gradients

Victor Gallego
Komorebi AI Technologies
victor.gallego@komorebi.ai

Abstract

This work proposes aesthetic gradients, a method to personalize a CLIP-conditioned diffusion model by guiding the generative process towards custom aesthetics defined by the user from a set of images. The approach is validated with qualitative and quantitative experiments, using the recent stable diffusion model and several aesthetically-filtered datasets. Code is released at https://github.com/vicgalle/stable-diffusion-aesthetic-gradients


Original SD
Personalized with SAC_{8+}
Original SD
Personalized with SAC_{8+}
Original SD
Personalized with $LAION_{7+}$

A fountain, sculpture

A giant octopus, bioluminescence

A nighttime cityscape, concept art

A gateway between dreams

A clearing filled with colorful plants in a thick wood where time has stopped, trending on Artstation

A beautiful mannequin made of marble printed in 3d geometric neon + kintsugi...

Photorealistic white marble greek goddess face profile sculpture entwined by crimson vines...

Marble Polished Tile. Dark Volcano is an impressive dark red quartzite...

A whole train covered in wildstyle graffiti in the style of Seen, HQ masterpiece, vibrant

Art Deco in Santorini island palm trees crystal clear neon water, trending Artstation, octane render...

A gothic cathedral in the middle of the forest

A painting of a path going up a hill, oil on canvas

Figure 1: Stable diffusion generations for the original model and personalized variants using SAC_{8+} and $LAION_{7+}$ aesthetic embeddings.

Introduction Recently developed text-to-image models have demonstrated an unprecedented capability to reason over natural language descriptions [8]. Their use, however, is constrained by the user's ability to describe the desired target through text. But can generative models really grasp the aesthetic vibe preferred by its users? Recent works such as *textual inversion* [2] and *dreambooth* [9]

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

aim to provide user personalization to diffusion models, but rather focus on learning custom objects from few images. Instead, in this work we present an alternative approach for *personalization* of text-to-image diffusion models. Our goal is to guide the generative process towards custom aesthetics defined by the user, without restricting us to single objects, but rather aesthetic patterns specified by arbitrarily large sets of images.

Method At first, the user chooses a textual prompt y to guide the generation. This prompt is passed through the CLIP text encoder [7] to obtain a textual embedding $c = \text{CLIP}_{\theta,txt}(y)$. Conditioned on this representation c, the diffusion process generates a image that matches the prompt. We propose to modify the previous representation c by taking into account another embedding, representing the aesthetic preferences of the user. Let $\{x_i\}_{i=1}^{K}$ be a set of K images representing the aesthetic preference of an user. We define its corresponding *aesthetic embedding* e as the average of the visual embeddings of the previous images, that is, $e = \frac{1}{K}\sum_{i=1}^{K} \text{CLIP}_{\theta,vis}(x_i)$. Then we normalize the resulting vector to be of unitary norm. The similarity between the two embeddings, computed as the dot product $ce^{\intercal}$, can be used to measure the agreement between CLIP representation of the textual prompt and the preferences of the user. Thus, the previous expression can be used as a loss and we can perform gradient descent with respect to CLIP text encoder weights to drive the prompt representation towards the aesthetics of the user: $\theta' = \theta + \epsilon\, \nabla_{\theta}\text{CLIP}_{\theta,txt}(y)e^{\intercal}$, with ϵ being a user-defined step size. After a few iterations, we compute the new, personalized prompt representation, $c' = \text{CLIP}_{\theta',txt}(y)$, and the generation continues using the underlying diffusion process. The resulting representation c' is more aligned to the user preference, while preserving the original semantics, as we will see in the experiments from the next section. Note that only the weights of the CLIP text encoder are modified, nor the visual encoder nor any other component of the diffusion model.

The benefits of our approach to personalization are several: (i) it works agnostically of the diffusion model, that is, it only requires a diffusion model which conditions on a textual prompt processed by CLIP. In the experiments we use the recent *stable diffusion* (SD) model. (ii) it is computationally cheap, as it only requires a few gradient steps (less than 20 in the experiments) of the CLIP text encoder. It is not necessary to fine-tune the diffusion model, thus making it GPU-friendly. (iii) the user only needs to store one aesthetic embedding per set of images, thus saving storage space and being amenable to sharing. In the case of CLIP-L/14, the variant used by SD, e is a vector of 768 dimensions. A potential drawback of our approach is that we introduce two new hyperparameters: the step size ϵ and the number of iterations. For our experiments, we set $\epsilon = 1e-4$ and vary the number of iterations from 5 to 20; yet these are two levers that the user can control.

Experiments and results First, we qualitatively show the effectiveness of our aesthetic gradients approach using several aesthetic embeddings and a collection of prompts of varying length and complexity. As the aesthetic embeddings, for this experiment we use two sets of images: SAC_{8+}, a subset of Simulacra Aesthetic Captions [6] with images filtered to have a rating greater than 8; and LAION_{7+}: a subset of LAION Aesthetics v1 [4] with images filtered to have a rating greater than 7. Figure 1 shows several generations, comparing the original SD model with the personalizations. Note that the personalized generations better reflect the aesthetics of each embedding, such as more fantasy-like imagery in the case of SAC_{8+}, and more floral patterns in LAION_{7+}.

We also perform an experiment to quantitatively assess the performance of the aesthetic gradient approach. To do so, we generate a list of 25 prompts, of varying length and complexity (Table 1). For each, we generate six images, both using the original SD model and personalizing with SAC_{8+}. Finally, for each image we compute its aesthetic score using an open source model [1]. The distribution of the scores for each of the two groups of generations is shown in Figure 2. Note the personalized models has improved aesthetic score, even though the diffusion model has not been optimized with respect to it. Appendix includes another batch of experiments, using different aesthetics.

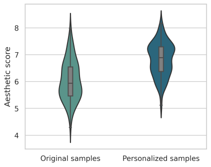


Figure 2: Aesthetic scores distribution

Conclusion We have proposed a flexible and efficient approach for personalizing text-to-image models, in particular by guiding the generation towards the aesthetic preferences of the user. As further work, it is straightforward to adapt our method to CLIP-guided diffusion [5], in which the customized CLIP model guides the generation at every timestep of the diffusion process. Also, instead of gradient descent optimizers, we could use SG-MCMC samplers [10, 3] to explore a greater region of the latent space and improve the diversity of the results.

Ethical Implications

This work aims to provide users with an effective framework for adapting the output of diffusion models to the preferred aesthetics of the user. While general text-to-image models might be biased towards specific attributes when synthesizing images from text, our approach enables the user to better reflect the desirable effects. On the other hand, malicious agents might try to use generated images to mislead viewers. This is a common issue, existing in other generative modeling approaches. Future research in generative modeling and personalization must continue investigating and revaluing these issues.

References

[1] Katherine Crowson. Simulacra aesthetic models. Technical Report Version 1.0, 2022. url https://github.com/crowsonkb/simulacra-aesthetic-models .

[2] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

[3] Victor Gallego and David Rios Insua. Stochastic gradient mcmc with repulsive forces. *arXiv preprint arXiv:1812.00071*, 2018.

[4] LAION. Laion aesthetics v1. Technical Report Version 1.0, LAION AI, 2022. url https://github.com/LAION-AI/aesthetic-predictor .

[5] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[6] John David Pressman, Katherine Crowson, and Simulacra Captions Contributors. Simulacra aesthetic captions. Technical Report Version 1.0, Stability AI, 2022. url https://github.com/JD-P/simulacra-aesthetic-captions .

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[9] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022.

[10] Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. *arXiv preprint arXiv:1902.03932*, 2019.

Supplementary material. Additional experiments and results

Figure 3 depicts further qualitative experiments, with different aesthetics of those from the main text. In particular, we target the following aesthetic preferences: `Aivazovsky` (five paintings from the artist), `cloudcore`, `gloomcore`, and `glowwave`, with the last three consisting in 100 random images scraped from Pinterest using those terms as the keyword. In the Figure, the first column shows the generations for the original SD model, the second column shows the generations also for the original model but appending the aesthetic keyword at the end of the prompt; and the third column show the generations personalized with the proposed aesthetic gradients approach. Note that whereas the second column shows that the effect of appending the aesthetic keyword in the prompt has a negligible or very dim effect due to the limitations of the CLIP text encoder, the aesthetic gradients approach can better reflect the aesthetic effect that is to be expected, dramatically improving the corresponding aesthetic vibe of the generations.

Table 1: Prompts used for the quantitative experiment

A fountain, sculpture
A pyramid over a snowy scenery
A giant octopus, bioluminescence
A still life of flowers, volumetric lighting
A still life of flowers, stained glass
A nighttime cityscape, concept art
The sacred library by Simon Stålenhag and Thomas Kinkade, oil on canvas
A gateway between dreams
Space jellyfish, watercolor
Giant skull without a lower jaw, floating above a pile of gems while it leaks gems and bone. An orange, cloudy sky fills the background
An orange overstuffed chair, custom design
Ethereal
A clearing filled with colorful plants in a thick woods where time has stopped, trending on Artstation
An archer lounging against a tree with petals falling, painting by Horace Vernet
Textless, 8k, hyperdetail Papier-mache, Ambient occlusion High key light, Contour rivalry, octane render redshift render, Porcelain painted ceramics by Krystle Mitchell, The efficient panda surrounds bangle, ascot plain peel postfix circadian sunroom
Dimming dares to swifting ruins lights, charges changes on the skies from above, blinks true to throughout, to a closing of hands on spacing world to binding breaks of recreating strings, then to dusting fantasy of hands that try to wave a reach of each, and a spine of splitting reeks of falling sense of decaying skying
Centuries of citadels, and been tuning in tones that been crystalize in a field that felt a widing in its own, still a lighten abyss vision for depths, it still crystalline in souls that truly enjoyed, of a meaning
White marble, white marble bas relief profile sculpture of a beautiful black haired woman with pale skin and a crown on her head sitted on an intricate metal throne, medusa, white and gold kintsugi, feminine shapes, crabs, spiders, scorpions, tarantulas, stunning, art by hr geiger and ridley scott and alphonse mucha and josephine wall, highly detailed, intricately detailed
Photorealistic white marble greek goddess face profile sculpture entwined by golden and crimson vines and roots, flesh shows at some parts under the broken marble, swirling liquified meat and red kintsugi, symbolist, visionary, etheric, entwined with iridiscent fractal lace, alien botanicals, cinematic composition, cinematic lighting
A beautiful mannequin made of marble printed in 3 d geometric neon + kintsugi, facing a giant doorway opening with a neon pink light, flowering iridescent pineapples + orchids, transcendent, vibrant color, clean linework, finely detailed, 4k, trending on artstation, photorealistic, volumetric lighting, octane render
A pirate ship, sepia coloring, hyper-detailed, dusk, 4k octane render
Vaporwave soviet skyline at sunrise, trending on Artstation. Many intrincate details
Marble Polished Tile. Sky Blue is an impressive pale blue quartzite. Its appearance is reminiscent of a splendid blue sky interspersed with fluffy white clouds, as its name suggests. This natural stone's base shuffles different soft blues such as blue lavender, pale blue, and pastel indigo. The veins look like clouds. Decorative marble tile
A photograph of an astronaut riding a horse
A painting of a tree, oil on canvas

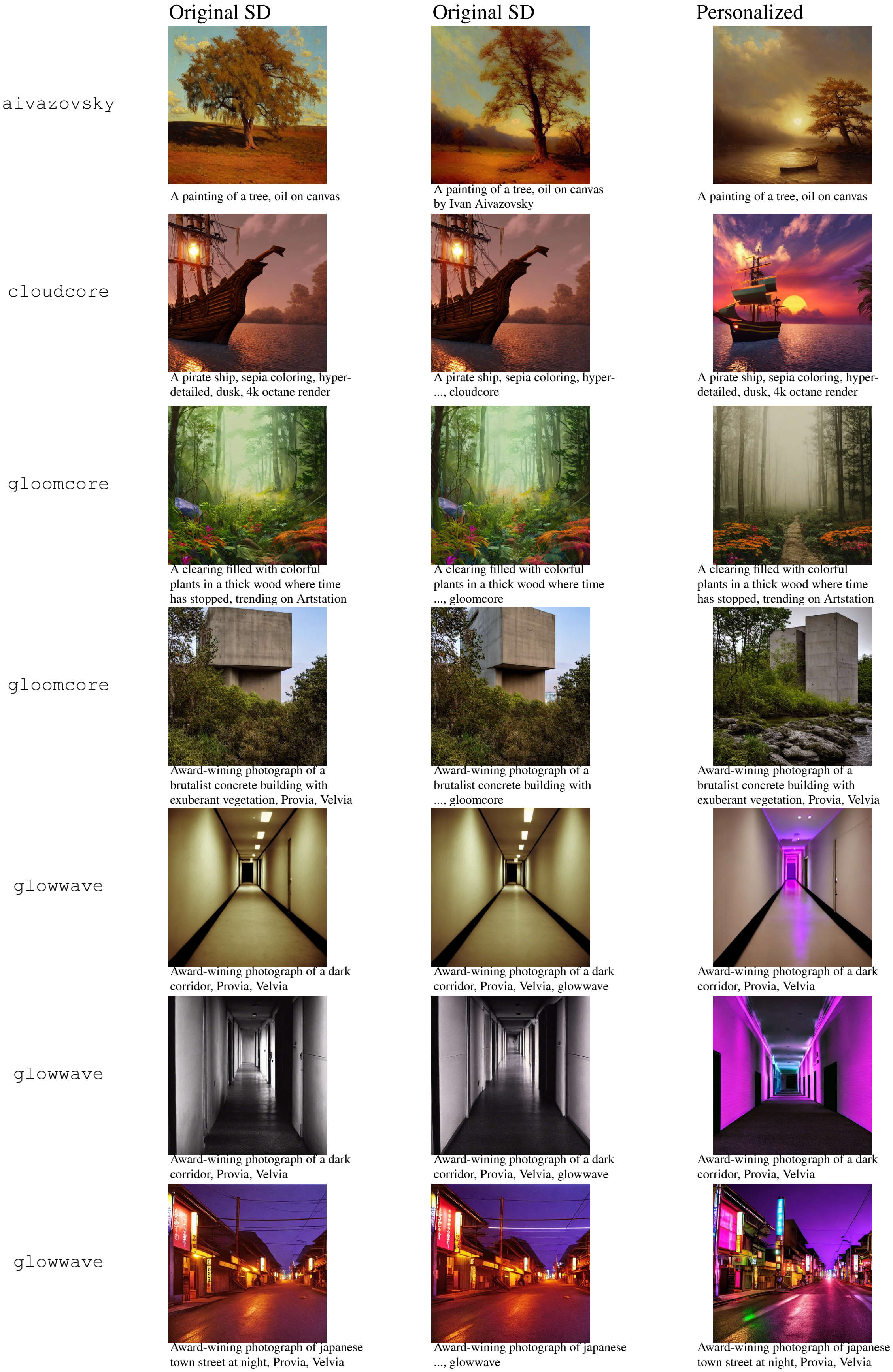

Original SD
Original SD
Personalized
aivazovsky
A painting of a tree, oil on canvas
A painting of a tree, oil on canvas by Ivan Aivazovsky
A painting of a tree, oil on canvas
cloudcore
A pirate ship, sepia coloring, hyper-detailed, dusk, 4k octane render
A pirate ship, sepia coloring, hyper-..., cloudcore
A pirate ship, sepia coloring, hyper-detailed, dusk, 4k octane render
gloomcore
A clearing filled with colorful plants in a thick wood where time has stopped, trending on Artstation
A clearing filled with colorful plants in a thick wood where time ..., gloomcore
A clearing filled with colorful plants in a thick wood where time has stopped, trending on Artstation
gloomcore
Award-wining photograph of a brutalist concrete building with exuberant vegetation, Provia, Velvia
Award-wining photograph of a brutalist concrete building with ..., gloomcore
Award-wining photograph of a brutalist concrete building with exuberant vegetation, Provia, Velvia
glowwave
Award-wining photograph of a dark corridor, Provia, Velvia
Award-wining photograph of a dark corridor, Provia, Velvia, glowwave
Award-wining photograph of a dark corridor, Provia, Velvia
glowwave
Award-wining photograph of a dark corridor, Provia, Velvia
Award-wining photograph of a dark corridor, Provia, Velvia, glowwave
Award-wining photograph of a dark corridor, Provia, Velvia
glowwave
Award-wining photograph of japanese town street at night, Provia, Velvia
Award-wining photograph of japanese ..., glowwave
Award-wining photograph of japanese town street at night, Provia, Velvia

Figure 3: Further qualitative results using different aesthetic embeddings.