

The Stable Artist: Steering Semantics in Diffusion Latent Space

Manuel Brack¹Patrick Schramowski^{1,3,4,5}Felix Friedrich^{1,3}Dominik Hintersdorf¹Kristian Kersting^{1,2,3,4}¹Computer Science Department, TU Darmstadt²Centre for Cognitive Science, TU Darmstadt, ³Hessian Center for AI (hessian.AI)⁴German Research Center for Artificial Intelligence (DFKI), ⁵LAION

{lastname}@cs.tu-darmstadt.de



Figure 1. The Stable Artist eases the generation of images via Stable Diffusion by (iterative) guidance. Original image (left images) generated using the prompt on top of image pair. Guidance prompt (bottom of image pair) and result (right). (Best viewed in color)

Abstract

Large-scale, text-conditioned generative diffusion models have recently drawn a lot of interest for their astonishing ability to produce high-fidelity images from text only. However, generating high-quality images in a single shot is nearly impossible and typically requires several small changes to the input prompt. Unfortunately, small changes to the input prompt often result in completely different images being generated, leaving the artist with little control.

We present the Stable Artist to enable control by allowing the artist to steer the diffusion process along a variable number of semantic directions. This semantic guidance (SEGA) allows for subtle edits to images, changes in composition and style, as well as optimization of the overall

artistic conception. Furthermore, SEGA enables probing latent spaces to gain insights into the representation of concepts learned by the model, even complex ones such as ‘carbon emission’. We demonstrate the Stable Artist on several tasks, showcasing high-quality image editing and composition.

1. Introduction

Generative AI methods are improving rapidly, and using text-to-image diffusion models (DM) [10–12] makes it now possible to generate text and images simply based on text input, producing impressive results on generative image tasks. Unfortunately, however, unraveling the concepts they learn during training and understanding how to influ-

ence what they actually output remains an open question. Users need deep knowledge of style-specific, long, and obfuscated prompts that usually eludes non-experts. Hence, high-quality images are rarely the result of the initially generated output, making a one-shot text-to-image generation infeasible. The contrary is the case; text-guided image generation is a highly iterative process that requires interaction between the model and its user. Consequently, the human user is likely to generate many images with slight changes to the text prompt and other parameters in order to achieve the envisioned outcome. In this ‘artist-in-the-loop’ setting, fine-grained control over the generated output and its elements is imperative but hardly feasible through current methods. Small changes in the phrasing of the prompt text may lead to the generation of entirely different images.

The required amount of control is generally only possible through providing image masks in combination with an edit instruction of the masked area. This is inherently limited, as it discards important structural information and the global composition of the image. Furthermore, some tasks like texture or style changes are not achievable with inpainting techniques. Other approaches like Prompt-to-Prompt (P2P) [3] rely on a form of soft, implicit masking by interacting with the attention masks of the input prompt. P2P utilizes the changes in attention maps between the original prompt and an edited prompt to target the relevant regions of the image. However, the granularity of control remains limited to the rather core-grained dimensions of the attention mask, and these approaches are inherently restricted to one editing operation at a time. On the other hand, Composable Diffusion [7] does enable conditioning on multiple concepts but only provides control over the initial image composition and does not support more subtle changes.

We present the Stable Artist, an iterative approach for guiding a generated image toward the desired output. The Stable Artist can change aspects of the initial image using Semantic Guidance (SEGA) that provides fine-grained control of the image generation process by leveraging sophisticated operations in the model’s latent space. This enables subtle edits to images, changes in composition and style, as well as optimization of the overall artistic conception. Furthermore, SEGA allows for probing the latent space of diffusion models to gain insights into how abstract concepts are represented by the model and how their interpretation reflects on the generated image. SEGA also facilitates advanced arithmetics between concepts that were previously only observed for natural language embeddings [1, 6, 8].

The Stable Artist supports editing with multiple concepts simultaneously while providing full control over the extent of changes to the image as well as the strength of each applied concept. All while using no masks—be it explicit or based on attention—and without any fine-tuning.

Algorithm 1 Semantic Guidance (SEGA)

```

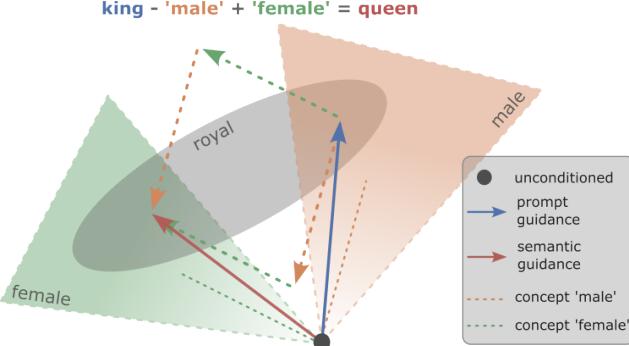
Require: model weights  $\theta$ , text condition  $text_p$ , edit texts List( $text_e$ ) and diffusion steps  $T$ 
Ensure:  $s_m \in [0, 1]$ ,  $\nu_{t=0} = 0$ ,  $\beta_m \in [0, 1]$ ,  $\lambda^i \in [-1, 1]$ ,  $s_e^i \in [0, 5000]$ ,  $\delta \in [0, 20]$ ,  $t = 0$ 
DM  $\leftarrow$  init-diffusion-model( $\theta$ )
 $c_p \leftarrow$  DM.encode( $text_p$ )
List( $c_e$ )  $\leftarrow$  DM.encode(List( $text_e$ ))
 $latents \leftarrow$  DM.sample( $seed$ )
while  $t \neq T$  do
     $n_\emptyset, n_p \leftarrow$  DM.predict-noise( $latents, c_p$ )
    List( $n_e$ )  $\leftarrow$  DM.predict-noise(List( $c_e$ ))  $\triangleright$  Eq. (6)
     $\mu_t \leftarrow 0$ 
    for all  $n_e$  in List( $n_e$ ) do
         $\phi_t^i \leftarrow s_e^i * (n_p - n_e)^{-1}$   $\triangleright$  Eq. (7)
        if positive guidance then
             $\mu_t^i \leftarrow$  where( $n_p - n_e > \lambda$ , max(1,  $|\phi_t|$ ))  $\triangleright$  Eq. (6)
             $\gamma_t^i \leftarrow \mu_t^i * (n_\emptyset - n_e)$   $\triangleright$  Eqs. (4) and (5)
        else
             $\mu_t^i \leftarrow$  where( $n_p - n_e < \lambda$ , max(1,  $|\phi_t|$ ))  $\triangleright$  Eq. (6)
             $\gamma_t^i \leftarrow \mu_t^i * (n_e - n_\emptyset)$   $\triangleright$  Eqs. (4) and (5)
        end if
    end for
     $\gamma_t \leftarrow \sum_{i \in I} g_i * \gamma_t^i$   $\triangleright$  Eq. (10)
     $\gamma_t \leftarrow \gamma_t + s_m * \nu_t$   $\triangleright$  Eq. (8)
     $\nu_{t+1} \leftarrow \beta_m * \nu_t (1 - \beta_m) * \gamma_t$   $\triangleright$  Eq. (9)
    if  $t \geq \delta$  then
         $pred \leftarrow s_g * (n_p - n_\emptyset - \gamma_t)$   $\triangleright$  Eq. (3)
    else
         $pred \leftarrow s_g * (n_p - n_\emptyset)$   $\triangleright$  Eq. (2)
    end if
     $latents \leftarrow$  DM.update-latents( $pred, latents$ )
     $t \leftarrow t + 1$ 
end while
 $image \leftarrow$  DM.decode( $latents$ )

```

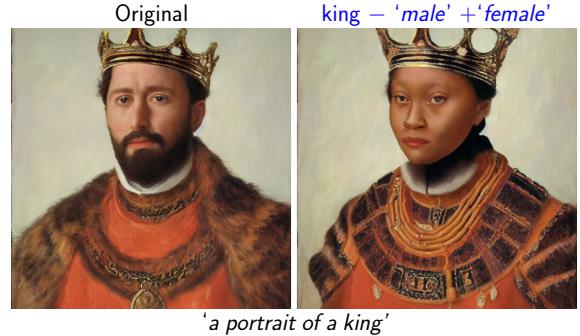
2. The Stable Artist

Let us now devise the Stable Artist¹, a semantic image editing technique for latent diffusion models. Roughly speaking, it generalizes the generative diffusion process by combining text conditioning through classifier-free guidance with editing concepts targeting dedicated parts of the image. In doing so, it substantially extends the combination of latents introduced in composable diffusion [7] and is a general-purpose advancement of Safe Latent Diffusion (SLD) [13]; while SLD suppresses one inappropriate direction during generation, the Stable Artist features a variable number of directions that can be suppressed or enforced.

¹In this context, ‘Stable’ is not only a nod to Stable Diffusion that builds the base of our implementation. Additionally, the Stable Artist behaves stable with respect to its control over the generated image.



(a) A (latent) diffusion process inherently organizes concepts and learns implicitly relationships between them, although there is no supervision.



(b) Guidance arithmetic: Guiding the image ‘a portrait of a king’ (left) using ‘king’ – ‘male’ + ‘female’ results in an images of a ‘queen’ (right).

Figure 2. Semantic guidance (SEGA) applied to the image ‘a portrait of a king’ using ‘king’ – ‘male’ + ‘female’’. (Best viewed in color)

2.1. Guided Diffusion

The first step towards the Stable Artist is guided diffusion. Specifically, diffusion models iteratively denoise a Gaussian distributed variable to produce samples of a learned data distribution. For text-to-image generation, the model is conditioned on a text prompt p and guided towards an image, faithful to that prompt. The training objective of a diffusion model \hat{x}_θ , can be written as

$$\mathbb{E}_{\mathbf{x}, \mathbf{c}_p, \epsilon, t} [w_t || \hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \omega_t \epsilon, \mathbf{c}_p) - \mathbf{x} ||_2^2] \quad (1)$$

where $(\mathbf{x}, \mathbf{c}_p)$ is conditioned on text prompt p , t is drawn from a uniform distribution $t \sim \mathcal{U}([0, 1])$, ϵ sampled from a Gaussian $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, and w_t, ω_t, α_t influence image fidelity depending on t . Consequently, the DM is trained to denoise $\mathbf{z}_t := \mathbf{x} + \epsilon$ to yield \mathbf{x} with the squared error as a loss. At inference, the DM is sampled using the model’s prediction of $\mathbf{x} = (\mathbf{z}_t - \bar{\epsilon}_\theta)$, with $\bar{\epsilon}_\theta$ as described below.

Classifier-free guidance [5] is a conditioning method using a purely generative diffusion model, eliminating the need for an additional pre-trained classifier. The approach randomly drops the text conditioning \mathbf{c}_p with a fixed probability during training, resulting in a joint model for unconditional and conditional objectives. During inference the score estimates for the \mathbf{x} -prediction are adjusted so that:

$$\tilde{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}_p) := \epsilon_\theta(\mathbf{z}_t) + s_g(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p) - \epsilon_\theta(\mathbf{z}_t)) \quad (2)$$

with guidance scale s_g and ϵ_θ defining the noise estimate with parameters θ . Intuitively, the unconditioned ϵ -prediction is pushed in the direction of the conditioned one, with the s_g determining the extent of the adjustment.

2.2. Semantic Guidance

Now, the main idea is to influence the diffusion process along several directions. To achieve this, SEGA substantially extends the principles introduced in classifier-free

guidance. The idea is to use multiple editing prompts e_i targeting arbitrary concepts of the generated image, in addition to the text prompt p .

One Direction. To introduce semantic guidance, let us start off with a single direction, i.e., editing prompt. Specifically, we use three ϵ -predictions with the goal of moving the unconditioned score estimate $\epsilon_\theta(\mathbf{z}_t)$ towards the prompt conditioned estimate $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p)$ and simultaneously away or also towards the concept conditioned estimate $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_e)$, depending on the editing direction. Formally, we compute

$$\begin{aligned} \bar{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_e) = \\ \epsilon_\theta(\mathbf{z}_t) + s_g(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p) - \epsilon_\theta(\mathbf{z}_t) - \gamma(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_e)) \end{aligned} \quad (3)$$

with the semantic guidance term γ

$$\gamma(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_e) = \mu(\mathbf{c}_p, \mathbf{c}_e; s_e, \lambda) \psi(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_e) \quad (4)$$

where μ applies an editing guidance scale s_e element-wise, and ψ depends on the editing direction:

$$\begin{aligned} \psi(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_e) = \\ \begin{cases} \epsilon_\theta(\mathbf{z}_t) - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_e) & \text{if pos. guidance} \\ -(\epsilon_\theta(\mathbf{z}_t) - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_e)) & \text{if neg. guidance} \end{cases} \end{aligned} \quad (5)$$

Consequently, changing the guidance direction is reflected by the direction of the vector between $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_e)$ and $\epsilon_\theta(\mathbf{z}_t)$. But back to μ . It considers those dimensions of the prompt conditioned estimate that are relevant to the defined editing prompt e . To this end, μ scales the element-wise difference between the prompt conditioned estimate and edit conditioned estimate by s_e for all elements where this difference

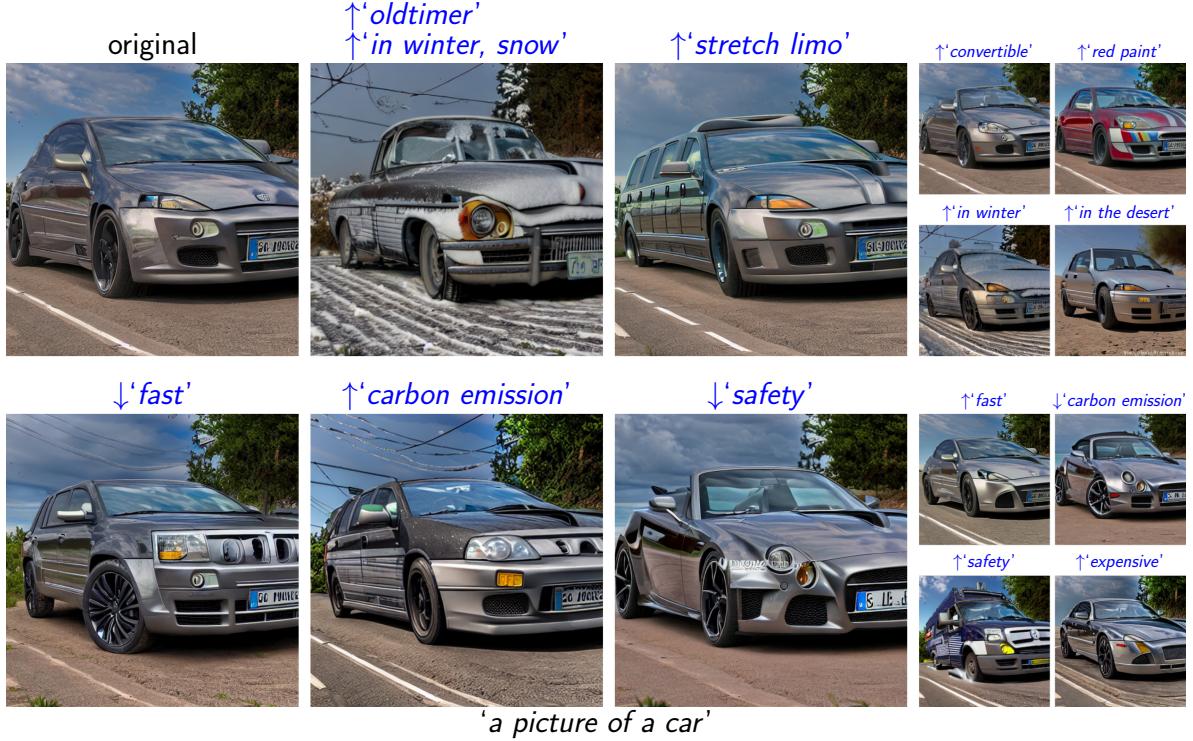


Figure 3. Image editing, performed using semantic guiding of the Stable Artist. All images generated from the same initial noise latent using the prompt ‘a picture of a car’. Editing prompts denoted in blue. Arrows indicate the editing direction. The Stable Artist can act on explicit edits for local and global changes of the image, as well as abstract editing concepts. (Best viewed in color)

is below or above a threshold λ and equals 0 otherwise:

$$\mu(\mathbf{c}_p, \mathbf{c}_e; s_e, \lambda) = \begin{cases} \max(1, |\phi|), & \text{where } \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p) \ominus \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_e) \leq \lambda \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

$$\text{with } \phi = s_e (\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p) - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_e))^{-1} \quad (7)$$

with both larger absolute values of λ and larger s_e leading to a more substantial shift away from the prompt text. For positive guidance, values greater than the threshold are considered and vice versa for negative guidance. Consequently, the former should choose negative values for λ and the latter positive ones.

Note that we clip the scaling factor of μ in order to avoid producing image artifacts. Following [4, 12], the values of each \mathbf{x} -prediction should adhere to the training bounds of $[-1, 1]$ to prevent low fidelity images.

To offer even more control over the diffusion process, we make two adjustments to the methodology presented above. We add a warm-up parameter δ that will only apply guidance γ after an initial warm-up period in the diffusion process, i.e., $\gamma(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_e) := \mathbf{0}$ if $t < \delta$. Naturally, higher values for δ lead to less significant adjustments of the generated image. If we aim to keep the overall composition of

the image unchanged, selecting a sufficiently high δ ensures that only fine-grained details of the output are altered.

Furthermore, we add a momentum term ν_t to the semantic guidance γ in order to accelerate guidance over time steps for dimensions that are continuously guided in the same direction. Hence, γ_t is defined as:

$$\gamma_t(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_e) = \mu(\mathbf{c}_p, \mathbf{c}_e; s_e, \lambda) \psi(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_e) + s_m \nu_t \quad (8)$$

with momentum scale $s_m \in [0, 1]$ and ν being updated as

$$\nu_{t+1} = \beta_m \nu_t + (1 - \beta_m) \gamma_t \quad (9)$$

where $\nu_0 = \mathbf{0}$ and $\beta_m \in [0, 1]$. Thus, larger β_m lead to less volatile changes of the momentum. Momentum is already built up during the warm-up period, even though γ_t is not applied during these steps.

Beyond One Direction. Now we are ready to move beyond using just one direction towards multiple concepts e_i and in turn combining multiple calculations of γ_t .

For all e_i , we calculate γ_t^i as described above with each defining their own parameter values λ^i, s_e^i . The adjusted $\hat{\gamma}_t$ is the result of the weighted sum of all γ_t^i :

$$\hat{\gamma}_t(\mathbf{z}_t, \mathbf{c}_p; \mathbf{e}) = \sum_{i \in I} g_i \gamma_t^i(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_{e_i}) \quad (10)$$

where g_i sums up to one. In order to account for different warm-up periods g_i is defined as $g_i = 0$ if $t < \delta_i$. However, momentum is built up using all edit prompt and applied once all warm-up periods are completed, i.e. $\forall \delta_i : \delta_i \geq t$. We provide a pseudo-code implementation of SEGA in Algorithm 1. Please note that this notation makes the simplified assumption of one single warm-up period δ for all edit prompts e_i .

2.3. Interactively Steering Semantics

Putting everything together, we start with an image generated from a prompt. Then, we iteratively perform adjustments with SEGA, enforcing or suppressing concepts, thus steering the overall semantics in diffusion latent space. Fig. 2 provides a 2-dimensional, visual explanation of semantic guidance. Intuitively, we can understand the latent space as a composition of arbitrary sub-spaces representing semantic concepts. The unconditioned noise estimate (black dot) starts us off at some random point in the latent space without semantic grounding. The guidance corresponding to the prompt “a portrait of a king” represents a vector (blue vector) moving us into a portion of the latent space where the concepts ‘male’ (human) and royal overlap, resulting in an image of a king. We can now further manipulate the generation process using SEGA. From the unconditioned starting point, we get the directions of ‘male’ and ‘female’ (orange/green lines) using estimates conditioned on the respective prompts. If we subtract this inferred ‘male’ direction from our prompt guidance and add the ‘female’ one, we now reach a point in latent space at the intersection of the ‘royal’ and ‘female’ sub-spaces, i.e. a queen. This vector represents the final direction (red vector) resulting from semantic guidance.

In the next section, we now illustrate the Stable Artist. The illustrations are based on our own implementation of Stable Diffusion version 1.5². The code, together with detailed demonstrations, is publicly available at github.com/ml-research/semantic-image-editing.

3. Interacting with Visual Concepts

In contrast to existing work, the Stable Artist directly interacts with concepts implicitly learned by DMs. Consider e.g. Fig. 2b. As one can see, the linear operations between noise estimates distilled using SEGA are semantically grounded. Arithmetic relations between concepts like these have previously been observed for natural language embeddings. In the depicted example, editing an image of a king by guiding it away from the concept ‘male’ and towards the concept ‘female’ produces a compositionally similar image but replaces the king with a queen. This indicates that the latent space of diffusion models may inherently be dis-

tangled to some extent, although further research in that direction is necessary to verify our assumption.

Furthermore, the Stable Artist can simultaneously change multiple aspects of an image with minimal interference of different concepts, as shown in Fig. 1 (top right) and Fig. 3 (top mid). We discuss the subtle control offered by SEGA in more detail in Sec. 4. Additionally, Fig. 3 highlights the versatile concepts that may be used to adjust images faithfully. These include concrete editing prompts that explicitly target certain portions of the image. In this regard, the stable artist supports both local edits to one of the depicted objects and global edits aimed at the environment of the image as a whole. Notably, multiple local and global edits may be combined arbitrarily. On the other hand, the Stable Artist also helps uncover how the underlying DM “interprets” more complex concepts and gives further insight into learned representations. For example, adding the concept ‘carbon emissions’ to the generated car produces a seemingly much older vehicle with a presumably larger carbon footprint. Similarly, reducing ‘safety’ yields a convertible with no roof and likely increased horsepower. Both these interpretations of the provided concepts with respect to cars are logically sound and provide valuable insights into the learned concepts of DMs. These experiments suggest a deeper natural language and image “understanding” that go beyond descriptive captions of images.

4. Fine-grained Image Editing & Composition

One of our driving forces is the vision that generative image creation and composition techniques should offer fine-grained control over the generated image. SEGA targets the relevant portions of the image while not changing other aspects. Furthermore, the Stable Artist offers continuous control over the strength of the applied concept semantically reflected in the image. It is worth noting that we achieve this without masking the image, be it user-provided masks or those implicitly inferred from attention maps. Instead, we produce results of competitive quality by targeting the relevant latent dimensions alone.

This level of control can be seen throughout the images shown in the present work. Considering the top right example in Fig. 1, we make two local changes to the image composition and a global style change affecting the entire image. All three operations are performed simultaneously in the same latent space, and nonetheless, the majority of the image remains unchanged down to small details like the structure of clouds, the bush in the foreground, the castle, or the hill it stands on. We can observe a similar level of control in Fig. 3, where the fore- and background remain largely unaffected by changes to the car and vice versa.

We investigated this further in Fig. 4, together with possible linear continuity of image alterations. Starting from the original image in the top left corner, we remove one

²<https://huggingface.co/runwayml/stable-diffusion-v1-5>

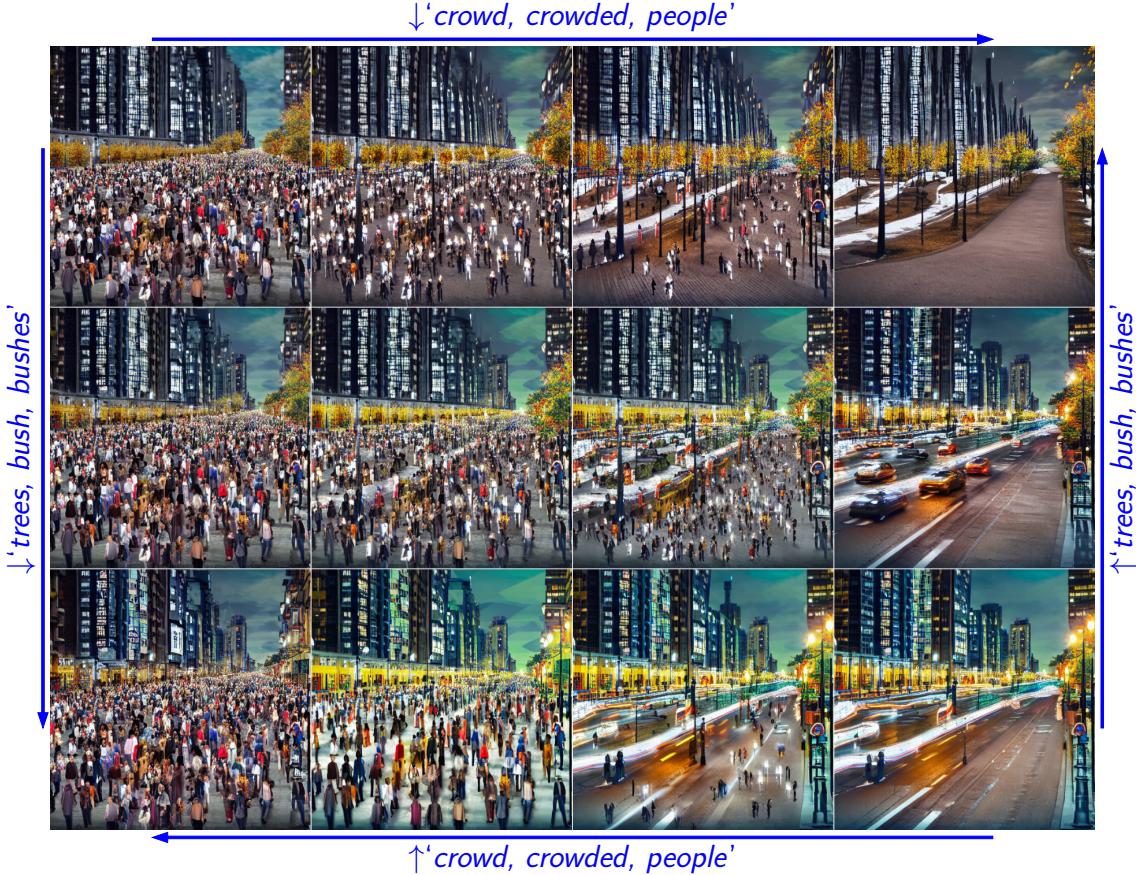


Figure 4. The Stable Artist offers strong control over the latent space and can gradually perform edits at the desired strength. All images generated from the same initial noise latent using the prompt ‘a crowded boulevard’. Editing prompts denoted in blue and are gradually increased in strength from left to right and top to bottom. (Best viewed in color)

concept from the image following the x- or y-axis. In both cases, the Stable Artist continually reduces the number of people or trees until the respective concept is removed entirely. Again, the rest of the image remains fairly consistent, especially the concept that is not targeted. Going even further, a similar level of control is also possible with an arbitrary mixture of applied concepts. Let us consider the second row of Fig. 4, for example. The number of trees is kept at a medium level in that the row of trees on the left side of each image is always removed, but the larger tree on the right remains. While keeping the number of trees stable, we can still gradually remove the crowd at the same time.

5. Latent Image Optimization

SEGA can be used to influence the overall style of an image as well as its quality with respect to certain artistic conceptions. Fig. 5 demonstrates examples of style transfer in which a photorealistic image is adjusted to reflect a different style of photography or painting. The Stable Artist faithfully

replicates the styles of well-known artists, as well as more abstract instructions. Again, the overall image composition remains largely unchanged, enabling a real transfer of the original image to the target style. This sets SEGA clearly apart from more simplistic techniques that only append the style instruction to the original prompt.

With Stable Artist, we can go even further than vanilla style transfer and actually optimize the overall look and quality of the generated output for arbitrary types of images directly in latent space. Creating high-fidelity images often requires a large amount of error-prone prompt engineering for every specific type of image and, in turn, results in long and obfuscated prompts. Instead, we can use a short and concise prompt and directly optimize the style through semantic guidance. To that end, we collected prompts known to produce high-quality results³ for five different types of images in *portrait photography, animation,*

³Prompts taken from <https://mpost.io/best-100-stable-diffusion-prompts-the-most-beautiful-ai-text-to-image-prompts/>

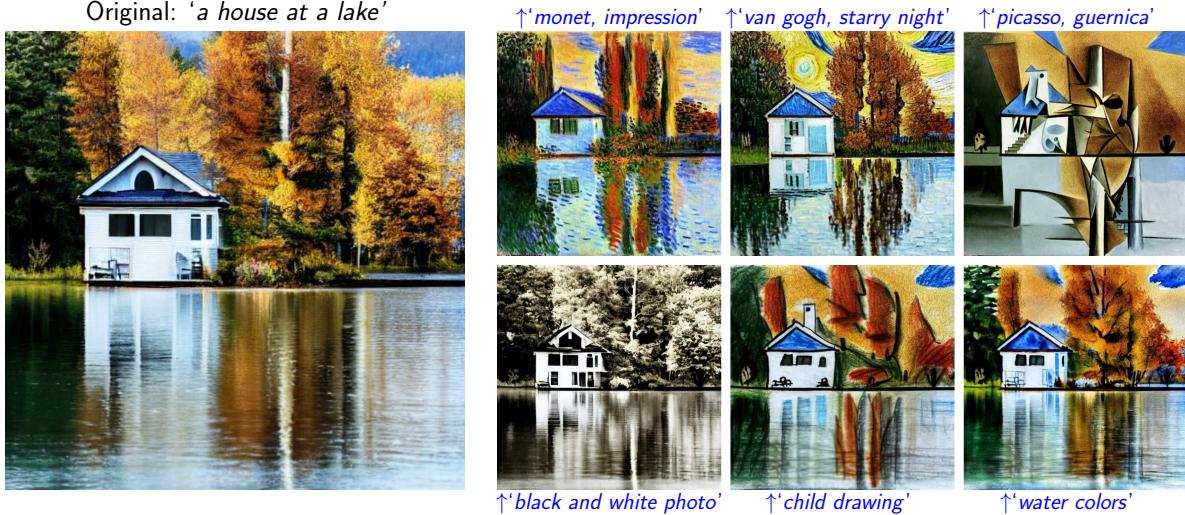


Figure 5. Style transfer performed by the Stable Artist. All images generated from the same initial noise latent using the prompt ‘a house at a lake’. Editing prompts denoted in blue. Arrows indicate the editing direction. (Best viewed in color)

concept art, character design, and modern architecture. We calculated the text embeddings for a set of prompts and took the average embedding per category as guidance conditioning. Exemplary results are depicted in Fig. 6. The results are of high quality and stay close to the original image, but accurately reflect the targeted artistic direction.

6. Qualitative Comparison

To investigate whether the Stable Artist provides more control to the artist, we compared it qualitatively to related techniques for text-to-image diffusion. While the fine-grained control and probing of the latent diffusion space are unique to SEGA, closest to the Stable Artist are Composable Diffusion [7] and Prompt-to-Prompt [3]. To compare them, we assess the performance of each approach on five different tasks depicted in Fig. 7. These consist of additive and subtractive image composition, style transfer, and the combination of composition and style changes.

All three methods perform well in adding elements to the image (Fig. 7 top left). Nonetheless, the Stable Artist is the only one keeping the riverside brick stone wall from the original image and also makes the least visible alterations to the castle and river. In the case of removing elements (Fig. 7 top right), Composable Diffusion alters large portions of the image composition, whereas P2P and the Stable Artist stay more faithful to the original image. This experiment further supports an interesting observation of the Stable Diffusion latent space. The model tends to prefer covering up image components targeted for removal with new details instead of generating the material that would be revealed behind or beneath the removed contents.

On the style transfer task (Fig. 7 bottom mid), Compos-

able Diffusion is incapable of combining the original image composition with the target style. For progressively stronger conditioning on van Gogh’s starry night, the approach simply shifts towards replicating the original painting and loses the image composition of the castle. However, P2P and the Stable Artist both create a faithful fusion between the content’s starting image and the target style.

When combining editing and style transfer (Fig. 7 bottom left & right), Composable Diffusion may generate images that do not fulfill any of the input descriptions. SEGA and P2P produce satisfactory results; however, the latter is limited to one editing prompt. Consequently, we generated these results by supplying both editing tasks in the same prompt for P2P. This also results in artifacts like multiple boats being added, although only one was specified.

Overall, Stable Artist outperforms Composable Diffusion in general image editing and composition on a variety of tasks. We achieve better or at least equally good performance in targeted editing as P2P but natively support conditioning on multiple concepts. Furthermore, SEGA does not rely on token-based attention and is therefore applicable to any conditioning beyond natural language, such as conditioning on images, averaged textual embeddings or embeddings calculated through textual inversion. We discuss this further in Sec. 8.

7. Broader Impact on Society

Recent developments in text-to-image models [9, 10, 12] have the potential for far-reaching impact on society, both positive and negative, when deployed in applications such as image generation, image editing, or search engines. Previous research [2, 13] described many potential negative so-

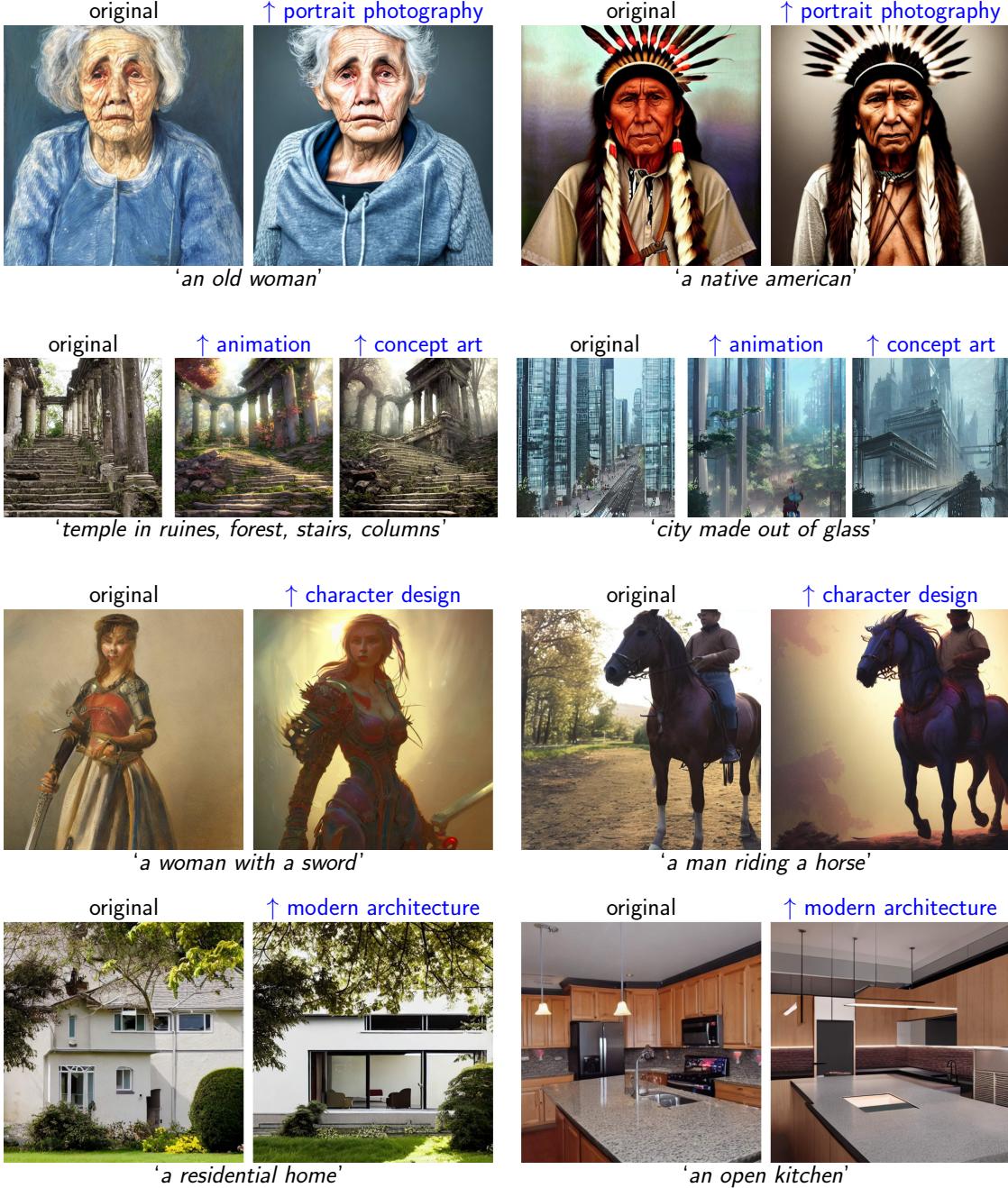


Figure 6. Latent style and image quality optimization using the Stable Artist. All image pairs/triplets are generated from the same initial noise latent and the text prompt under the images. Images are guided towards the average prompt embeddings of high-quality images belonging to the category stated in blue. (Best viewed in color)

cietal implications that may arise due to the careless use of such large-scale generative models. Many of these problems can be attributed to the noisy, large-scale datasets these models rely on. Since recent text-to-image models, such as stable diffusion, are trained on web-crawled data containing inappropriate content [14], they are no exception

to this issue. Specifically, current versions of stable diffusion show signs of inappropriate degeneration [13]. While Schramowski *et al.* utilize the model’s notion of inappropriateness to steer the model away from generating related content, it is noteworthy that we introduce an approach that could also be used to guide image generation toward inap-

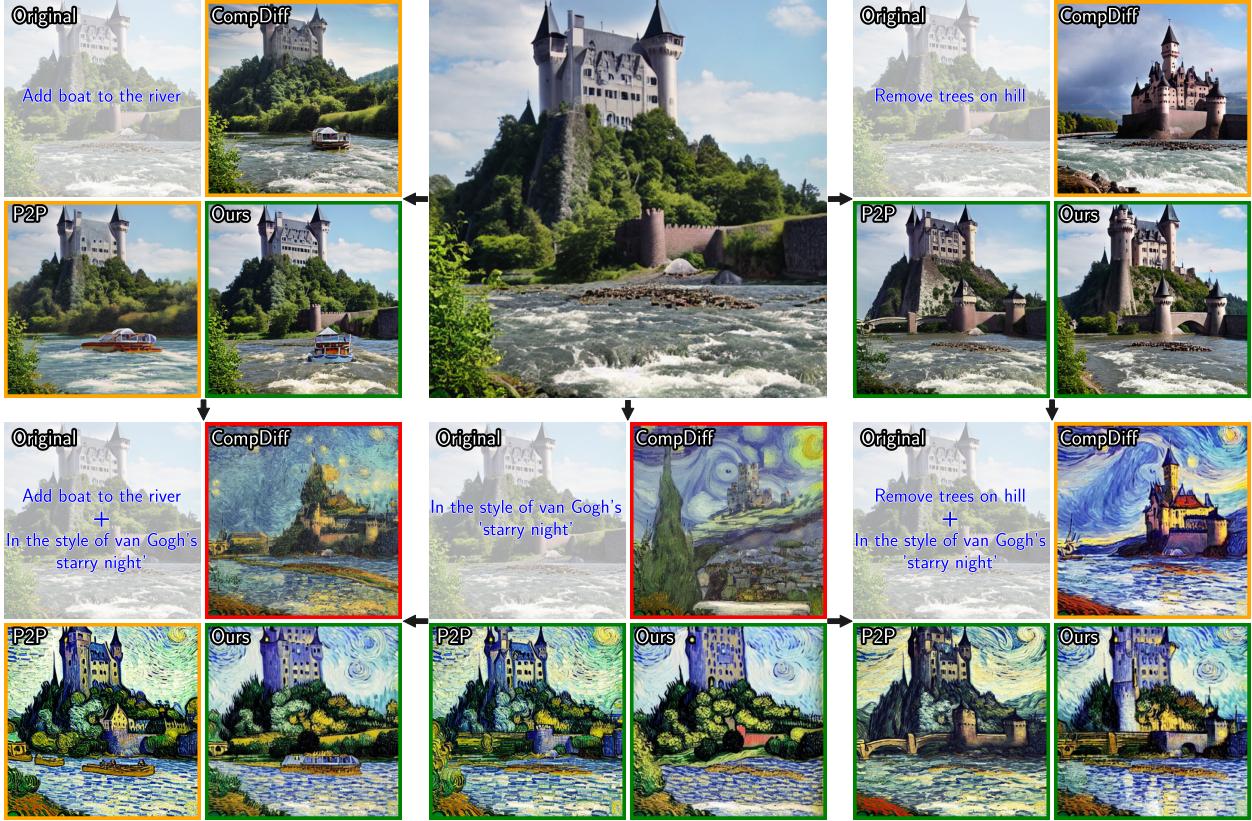


Figure 7. Comparison of image editing between Prompt-to-Prompt (P2P) [3], Composable Diffusion (CompDiff) [7], and the Stable Artist with SEGA (ours). All images were generated from the same initial noise latent using the prompt ‘a castle next to a river’. Colored boxes around the image denote the quality of the edit with respect to the editing task and the goal of minimal changes to the unaffected parts of the image. (Best viewed in color)

propriate material. However, on the positive side, SEGA could also have the potential to increase e.g. fairness, by detecting and steering certain related concepts. Therefore, we advocate for further research in this direction.

Another frequently voiced point of criticism, is the notion that generative models like stable diffusion are replacing human artists and illustrators. At first glance, the great results produced by these models might warrant this impression. Examples like the DALL-E generated Cosmopolitan cover⁴ subtitled with the phrase “And it only took 20 seconds to make” certainly seem to support this point. However, looking more closely at this example reveals that creating this cover still involved multiple hours spent by a human user interacting with the model. Consequently, we argue that generative models remain a tool to be used by humans for creating artwork. The generative process as a whole still requires a substantial amount of iterative human feedback and creative thinking. The introduced Stable Artist increases the interaction capabilities in these processes.

⁴<https://www.cosmopolitan.com/lifestyle/a40314356/dall-e-2-artificial-intelligence-cover/>

8. Conclusions

We presented the Stable Artist for directly interacting with concepts in DM’s latent space. To this end, we introduced semantic guidance (SEGA) that allows one to influence/steer the diffusion process along several directions. We demonstrated that the Stable Artist using SEGA offers fine-grained control over the generated image for performing sophisticated image composition and editing.

The Stable Artist covers several exciting avenues for future work. For instance, one should investigate more closely how concepts are represented in the latent space of DM’s and how to target and quantify them. More importantly, automatically detecting concepts could provide novel insights and toolsets to mitigate biases, as well as enacting privacy concerns of real people memorized by the model.

Acknowledgments. This research has benefited from the Hessian Ministry of Higher Education, Research, Science and the Arts (HMWK) cluster projects “The Third Wave of AI” and hessian.AI, from the German Center for Artificial

Intelligence (DFKI) project “SAINT”, the Federal Ministry of Education and Research (BMBF) project KISTRA (reference no. 13N15343), as well as from the joint ATHENE project of the HMWK and the BMBF “AVSV”.

References

- [1] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, 2012. [2](#)
- [2] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *CoRR*, abs/2211.03759, 2022. [7](#)
- [3] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. Preprint at <https://arxiv.org/abs/2208.01626>, 2022. [2, 7, 9](#)
- [4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Proceedings of the Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020. [4](#)
- [5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. [3](#)
- [6] Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, 2014. [2](#)
- [7] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. [2, 7, 9](#)
- [8] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013. [2](#)
- [9] Alexander Quinn 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. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2022. [7](#)
- [10] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. Preprint at <https://arxiv.org/abs/2204.06125>, 2022. [1, 7](#)
- [11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. [1](#)
- [12] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamvar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *CoRR*, abs/2205.11487, 2022. [1, 4, 7](#)
- [13] Patrick Schramowski, Manuel Brack, Björn Deisereth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. *arXiv preprint arXiv:2211.05105*, 2022. [2, 7, 8](#)
- [14] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. [8](#)