High-Fidelity Guided Image Synthesis with Latent Diffusion Models

Jaskirat Singh[1] Stephen Gould[1,2] Liang Zheng[1,2]
[1]The Australian National University [2]Australian Centre for Robotic Vision
{jaskirat.singh, stephen.gould, liang.zheng}@anu.edu.au

Figure 1. ***Overview***. We propose a novel stroke based guided image synthesis framework which *(Left)* resolves the intrinsic domain shift problem in prior works (b), wherein the final images lack details and often resemble simplistic representations of the target domain (e) (generated using only text-conditioning). Iteratively reperforming the guided synthesis with the generated outputs (c) seems to improve realism but it is expensive and the generated outputs might lose faithfulness with the reference (a) with each iteration. *(Right)* Additionally, the user is also able to specify the semantics of different painted regions without requiring any additional training or finetuning.

Abstract

Controllable image synthesis with user scribbles has gained huge public interest with the recent advent of text-conditioned latent diffusion models. The user scribbles control the color composition while the text prompt provides control over the overall image semantics. However, we note that prior works in this direction suffer from an intrinsic domain shift problem wherein the generated outputs often lack details and resemble simplistic representations of the target domain. In this paper, we propose a novel guided image synthesis framework, which addresses this problem by modelling the output image as the solution of a constrained optimization problem. We show that while computing an exact solution to the optimization is infeasible, an approximation of the same can be achieved while just requiring a single pass of the reverse diffusion process. Additionally, we show that by simply defining a cross-attention based correspondence between the input text tokens and the user stroke-painting, the user is also able to control the semantics of different painted regions without requiring any conditional training or finetuning. Human user study results show that the proposed approach outperforms the previous state-of-the-art by over 85.32% on the overall user satisfaction scores. Project page for our paper is available at https://1jsingh.github.io/gradop.

1. Introduction

Guided image synthesis with user scribbles has gained widespread public attention with the recent advent of large-scale language-image (LLI) models [23, 26, 28, 30, 40]. A novice user can gain significant control over the final image contents by combining text-based conditioning with unsupervised guidance from a reference image (usually a coarse stroke painting). The text prompt controls the overall image semantics, while the provided coarse stroke painting allows the user to define the color composition in the output scene.

Existing methods often aim attempt to achieve this through two means. The first category leverages conditional training using semantic segmentation maps [8, 28, 39]. However, the conditional training itself is quite time-consuming and requires a large scale collection of dense semantic segmentation labels across diverse data modalities. The second category, typically leverages an inversion based approach for mapping the input stroke painting to the target data manifold without requiring any paired annotations. For instance, a popular solution by [22, 35] introduces the painting based generative prior by considering a noisy version of the original image as the start of the reverse diffusion process. However, the use of an inversion based approach causes an intrinsic domain shift problem if the domain gap between the provided stroke painting and the target domain is too high. In particular, we observe that the resulting outputs often lack details and resemble simplistic representations of the target domain. For instance, in Fig. 1, we notice that while the target domain consists of *realistic photos* of a landscape, the generated outputs resemble simple pictorial arts which are not very realistic. Iteratively reperforming the guided synthesis with the generated outputs [4] seems to improve realism but it is costly, some blurry details still persist (refer Fig. 4), and the generated outputs tend to lose faithfulness to the reference with each successive iteration.

To address this, we propose a diffusion-based guided image synthesis framework which models the output image as the solution of a constrained optimization problem (Sec. 3). Given a reference painting y, the constrained optimization is posed so as to find a solution x with two constraints: 1) upon painting x with an autonomous painting function we should recover a painting similar to reference y, and, 2) the output x should lie in the target data subspace defined by the text prompt (*i.e.*, if the prompt says *"photo"* then we want the output images to be realistic photos instead of cartoon-like representations of the same concept). Subsequently, we show that while the computation of an exact solution for this optimization is infeasible, a practical approximation of the same can be achieved through simple gradient descent.

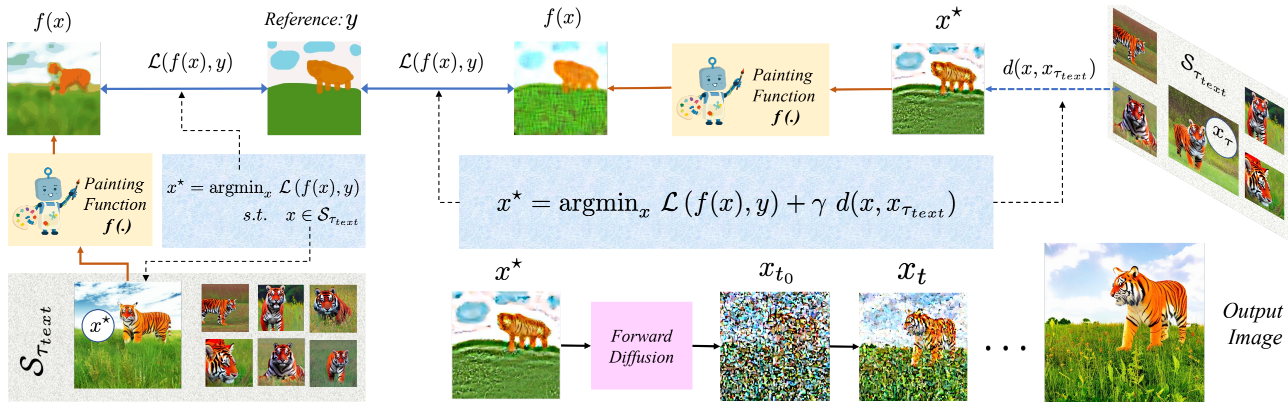
Finally, while the proposed optimization allows the user to generate image outputs with high realism and faithfulness (with reference y), the fine-grain semantics of different painting regions are inferred implicitly by the diffusion model. Such inference is typically dependent on the generative priors learned by the diffusion model, and might not accurately reflect the user's intent in drawing a particular region. For instance, in Fig. 1, we see that the light blue regions can be inferred as blue-green grass instead of a river. To address this, we show that by simply defining a cross-attention based correspondence between the input text tokens and user stroke-painting, the user can control semantics of different painted regions without requiring semantic-segmentation based conditional training or finetuning.

2. Related Work

GAN-based methods have been extensively explored for performing guided image synthesis from coarse user scribbles. [1–3, 14, 27, 37, 42] use GAN-inversion for projecting user scribbles on to manifold of real images. While good for performing small scale inferences these methods fail to generate highly photorealistic outputs when the given stroke image is too far from the real image manifold. Conditional GANs [7, 13, 18, 21, 24, 36, 43] learn to directly generate realistic outputs based on user-editable semantic segmentation maps. In another work, Singh *et al*. [34] propose an image synthesis framework which leverages autonomous painting agents [19, 32, 33, 44] for inferring photorealistic outputs from rudimentary user scribbles. Despite its efficacy, this requires the creation of a new dataset and conditional training for each target domain, which is expensive.

Guided image synthesis with LLI models [6, 23, 26, 28, 30, 40, 41] has gained widespread attention [9, 12, 15, 17, 20, 29, 31] due to their ability to perform high quality image generation from diverse target modalities. Of particular interest are works wherein the guidance is provided using a coarse stroke painting and the model learns to generate outputs conditioned on both text and painting. Current works in this direction typically 1) use semantic segmentation based conditional training [8, 28, 39] which is expensive, or, 2) adopt an inversion-based approach for mapping the input stroke painting to the target data manifold without requiring paired annotations. For instance, Meng *et al*. [22] propose guided image synthesis framework, wherein the generative prior is introduced by simply considering a noisy version of the original sketch input as the start of the reverse diffusion process. Choi *et al*. [5] propose an iterative conditioning strategy wherein the intermediate diffusion outputs are successively refined to move towards the reference image. While effective, the use of an inversion-like approach causes an implicit domain shift problem, wherein the output images though faithful to the provided reference show blurry or less textured details. Iteratively reperforming guided synthesis with generated outputs [4] seems to improve realism but it is costly. In contrast, we show that it is possible to perform highly photorealistic image synthesis while just requiring a single reverse diffusion pass.

Cross-attention control. Recently, Hertz *et al.* [10] propose a prompt-to-prompt image editing approach with text-conditioned diffusion models. By constraining the cross-attention features of all non-targeted text tokens to remain



(a) Constrained Optimization Formulation

(b) Obtaining an Approximate Solution

Figure 2. ***Method Overview.*** (a) Given a reference painting y and text prompt τ_{text}, we formulate the guided synthesis output as the solution $x^\star$ of a constrained optimization problem with 2 properties: 1) $x^\star$ lies in the subspace $\mathcal{S}_{\tau_{text}}$ of outputs conditioned only on the text, and, 2) upon painting x we should recover reference painting y. While computing an exact solution of this optimization is infeasible, we show that an approximation can be obtained in (b). Here we first solve an unconstrained approximation of the original optimization to compute a point $x^\star$ which is close to a random sample $x_{\tau_{text}} \in \mathcal{S}_{\tau_{text}}$ in the latent space, while still being faithful to the reference y. This $x^\star$ is then mapped back to the target subspace $\mathcal{S}_{\tau_{text}}$ using the diffusion based inversion from [35] to get the final image output.

the same, they show that by only modifying the text prompt, it is possible to perform diverse image editing without changing the underlying structure of the original input image. In contrast, we use cross-attention control for from-scratch synthesis and show that by simply defining a cross-attention based correspondence between input text tokens and the user stroke-painting, it is possible to control and define the fine-grain semantics of different painted regions.

3. Our Method

Let $f : \mathcal{D}_{real} \rightarrow \mathcal{D}_{paint}$ be a function mapping a real input image x to its painted image $f(x)$. Then given a colored stroke image y and input text prompt τ_{text}, we formulate the computation of guided image synthesis output $x^\star$ as the solution to the following constrained optimization problem,

$$x^\star = \text{argmin}_x \; \mathcal{L}\,(f(x), y) \tag{1}$$

$$\text{subject to } x \in \mathcal{S}_{\tau_{text}} \tag{2}$$

where $\mathcal{L}\,(f(x), y)$ represents a distance measure between the painted output $f(x)$ of image x and the target painting y, while $\mathcal{S}_{\tau_{text}}$ represents the subspace of output images conditioned only on the text input.

In other words, by additionally conditioning on a stroke image y, we wish to find a solution $x^\star$ such that 1) the distance between the painted image of x and reference painting y is minimized, while at the same time ensuring 2) the final solution lies in the subspace of images conditioned only on the text prompt τ_{text}. For instance, if the text says *"a realistic photo of a tree"* then the use of stroke-based guidance should still produce a *"realistic photo"*, wherein the composition of the tree regions is controlled by the painting y.

3.1. GradOP: Obtaining an Approximate Solution

The optimization problem in Eq. 1 can be reformulated as an unconstrained optimization problem as,

$$x^\star = \text{argmin}_x \; \mathcal{L}\,(f(x), y) + \gamma \; d(x, \mathcal{S}_{\tau_{text}}), \tag{3}$$

where $d(x, \mathcal{S}_{\tau_{text}})$ represents a distance measure between x and subspace $\mathcal{S}_{\tau_{text}}$, and γ is a hyperparameter.

A cursory glance at the above formulation should make it evident that the computation of an exact solution is infeasible without first generating a large enough sample size for the $\mathcal{S}_{\tau_{text}}$ subspace, which will be quite time consuming.

To address this, we propose to obtain an approximate solution by estimating $d(x, \mathcal{S}_{\tau_{text}})$ through the distance of x from a single random sample $x_{\tau_{text}} \in \mathcal{S}_{\tau_{text}}$. Thus, we can approximate the optimization problem as follows,

$$x^\star = \text{argmin}_x \; \mathcal{L}\,(f(x), y) + \gamma \; d(x, x_{\tau_{text}}). \tag{4}$$

Assuming a latent diffusion model with decoder $\mathcal{D}$, we can rewrite the above above optimization in latent space as,

$$z^\star = \text{argmin}_z \; \mathcal{L}\,(f(\mathcal{D}(z)), y) + \gamma \; \|z - z_{\tau_{text}}\|_2. \tag{5}$$

where the image output $x^\star$ can be computed as $x^\star = \mathcal{D}(z^\star)$.

In order to solve the above optimization problem, we first use the diffusion model to sample $x_{\tau_{text}} \in \mathcal{S}_{\tau_{text}}$. Initializing $z = \mathcal{E}(x_{\tau_{text}})$, where $\mathcal{E}$ represents the encoder, we solve the above optimization using gradient descent (assuming f and $\mathcal{L}$ are differentiable). Finally, we note that the solution $x^\star = \mathcal{D}(z^\star)$ to the above approximation of the optimization problem might be non-photorealistic, as Eq. 5 has no explicit constraint for enforcing $x \in \mathcal{S}_{\tau_{text}}$ (Eq. 2). We therefore use the diffusion-based-inversion approach from [35] in order to map it to the target image subspace $\mathcal{S}_{\tau_{text}}$.

Algorithm 1 GradOP: Solution Approximation

Input: Stroke Painting y, text prompt τ_{text}
Require: Differentiable painting function f, differentiable distance measure $\mathcal{L}$, hyperparameter γ, t_0.

Sample $x_{\tau_{text}} \in \mathcal{S}_{\tau_{text}}$;
Initialize $z = z_{\tau_{text}} = \mathcal{E}(x_{\tau_{text}})$;
for $0 \leq i \leq M$ **do**
    $\mathcal{L}_{total} = \mathcal{L}\left(f(\mathcal{D}(z)), y\right) + \gamma \left\|z - z_{\tau_{text}}\right\|_2$;
    $z = z - \lambda \nabla_z \mathcal{L}_{total}$;
end for
$z_{t_0} = \textsc{ForwardDiff}(z^\star = z, 0 \rightarrow t_0)$;
$z = \textsc{ReverseDiff}(z_{t_0}, t_0 \rightarrow 0)$;
return $x_{out} = \mathcal{D}(z)$.

In other words, unlike prior works [5, 22] which directly perform an inversion-like operation on the reference painting y (high domain gap with the target subspace *e.g.* real images), GradOP first uses the unconstrained optimization from Eq. 5 to compute a point $x^\star$ which is close to target subspace in the latent space z, while still being faithful to the reference y. Since $x^\star$ is closer to target domain than y, the inversion operation leads to more realistic outputs (refer Fig. 4). Please refer Alg. 1 for the detailed implementation.

3.2. GradOP+ : Improving Sampling Efficiency

While the guided synthesis solution in Alg. 1 leads to high output realism, for each output it first requires the sampling of a text-only conditioned image $x_{\tau_{text}} \in \mathcal{S}_{\tau_{text}}$. To address this, we propose a modified guided image synthesis approach which allows for equally high quality outputs while requiring just a single reverse diffusion pass for each output. Our key insight is that a lot of information in $z^\star$ is discarded during the forward diffusion from $z^\star \rightarrow z_{t_0}$. Thus, instead of performing the optimization to first compute $z^\star$, we would like to directly optimize the intermediate latent states z_t by injecting the optimization gradients within the reverse diffusion process itself (refer Fig. 3).

In particular, at any timestep t during the reverse diffusion, we wish to introduce optimization gradients in order to solve the following optimization problem,

$$z_t^\star = \operatorname{argmin}_z \mathcal{L}\left(f(\mathcal{D}(z)), y\right) + \gamma \left\|z - z_t\right\|_2. \quad (6)$$

However, the introduction of gradients will cause $z_t^\star$ to not conform with the expected latent distribution at timestep t. We therefore pass it through the forward diffusion process in order to map it back to the expected latent variable distribution. Please refer Alg. 2 for the detailed implementation.

3.3. Controlling Semantics of Painted Regions

Finally, while the above approximate guided image synthesis algorithm allows for generation of image outputs with high *faithfulness* and *realism*, the semantics of different painted regions are inferred in an implicit manner. Such inference is typically based on the cross-attention priors (learned by the diffusion model) between the provided text tokens and the input painting throughout the reverse diffusion process. For instance, in the first example from Fig. 5, we note that for different outputs, the blue region can be inferred as a river, waterfall, or a valley. Also note that some painting regions might be entirely omitted (*e.g.* the brown strokes for the hut), if the model does not understand that the corresponding strokes indicate a distinct semantic entity *e.g.* a hut, small castle *etc*. Moreover, as shown in Fig. 5 such discrepancies persist even if the corresponding text tokens (*e.g.* a hut) are added to the textual prompt.

Our key motivation is that the when generated faithfully, the average attention maps across different cross-attention layers show high overlap with the target object segmentation during the initial to intermediate parts of the reverse

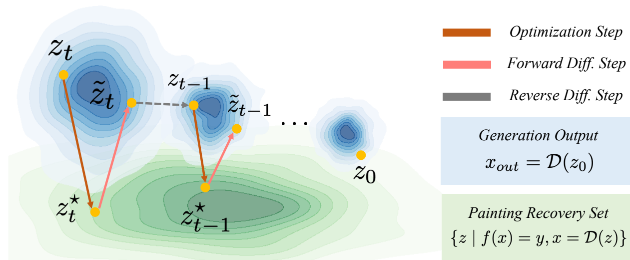


Figure 3. ***GradOP+ Overview.*** At any timestep t, the optimization in Eq. 6 ($z_t \rightarrow z_t^\star$) reduces the painting recovery loss, while the forward diffusion step $z_t^\star \rightarrow \tilde{z}_t$ maps it back to the expected latent distribution. By iteratively performing this optimization, *GradOP+* modifies the reverse sampling trajectory to lead to output $x_{out} = \mathcal{D}(z_0)$ which is also faithful to the target painting y.

Algorithm 2 GradOP+ : Improving Sampling Efficiency

Input: Stroke Painting y, text prompt τ_{text}
Require: Differentiable painting function f, distance measure $\mathcal{L}$, hyperparameter $\gamma, t_0, t_{start}, t_{end}$.

Sample $z_T \sim \mathcal{N}(0, \boldsymbol{I})$;
for $t = T-1, T-2 \ \ldots \ 0$ **do**
    $z_t = \textsc{ReverseDiff}(z_{t+1}, t+1 \rightarrow t)$;
    if $t_{start} \leq t \leq t_{end}$ **then**
        Initialize $z = z_t$;
        for $0 \leq i \leq M$ **do**
            $\mathcal{L}_{total} = \mathcal{L}\left(f(\mathcal{D}(z)), y\right) + \gamma \left\|z - z_t\right\|_2$;
            $z = z - \lambda \nabla_z \mathcal{L}_{total}$;
        end for
        $z_t = \textsc{ForwardDiff}(z_t^\star = z, 0 \rightarrow t)$
    end if
end for
return $x_{out} = \mathcal{D}(z_0)$.

diffusion process. In our experiments, we found the reverse to also be true. That is, by constraining the cross attention map corresponding to a target semantic label to have a high overlap with the desired painting region, it is possible to control the semantics of different painting regions without the need for segmentation based conditional training.

In particular, given the binary masks corresponding to different painting regions $\{\mathcal{B}_1, \dots \mathcal{B}_N\}$ and the corresponding semantic labels $\{u_1, \dots u_N\}$, we first modify the input text tokens as follows,

$$\tau_{modified} = \tau + \{\text{CLIP}(u_i) \mid i \in [1, N]\}\,, \qquad (7)$$

where τ is the set of CLIP [25] tokens for input text prompt.

At any timestep $t \in [0, T]$ during the reverse diffusion process, we then enforce semantic control by modifying the cross-attention map $\mathcal{A}_t^i$ corresponding to label u_i as follows,

$$\tilde{\mathcal{A}}_t^i = w_i \left[(1 - \kappa_t)\, \mathcal{A}_t^i + \kappa_t \, \frac{\mathcal{B}_i}{\|\mathcal{B}_i\|_F} \, \|\mathcal{A}_t^i\|_F \right] \qquad (8)$$

where $\|.\|_F$ represents the Frobenius norm, $\kappa_t = t/T \in [0, 1]$ helps regulate the overlap between the cross-attention output $\tilde{\mathcal{A}}_t^i$ and the desired painting region $\mathcal{B}_i$ during the reverse diffusion process, and, weights $w_i,\ i \in [1, N]$ help the user to control the relative importance of expressing different semantic concepts in the final image.

4. Experiments

Implementation Details. We use publicly available text-conditioned latent diffusion models [28, 38] for implementing the purposed approach in Sec. 3. The constrained optimization is performed using gradient descent with the Adam [16] optimizer and number of gradient steps $N_{grad} \in [20, 60]$ (please refer Sec. 5.2 for detailed analysis). While several formulations of the distance measure $\mathcal{L}$ and painting function f are possible (refer supp. material for details), we find that simply approximating the function $\mathcal{L}$ using mean squared distance and f as a convolution operation with a gaussian kernel seems to give the fastest inference time performance with our method. For consistency reasons, we use the non-differentiable painting function from SDEdit [22] while reporting quantitative results (refer Sec. 4.1).

4.1. Stroke Guided Image Synthesis

Evaluation metrics. Given an input stroke painting, we compare the performance of our approach with prior works in guided image synthesis when no paired data is available. The performance of the final outputs is measured in terms of both *faithfulness* of the generated image with the target stroke painting as well as the *realism* of the final output distribution. In particular, given an input painting y and output real image prediction x, we define faithfulness $\mathcal{F}(x, y)$ as,

$$\mathcal{F}(x, y) = \mathcal{L}_2(f(x), y) \qquad (9)$$

Method	Evaluation criteria		User Study Results	
	$\mathcal{F}(x, y) \downarrow$	$\mathcal{R}(.) \downarrow$	Realism $\uparrow$	Satisfaction $\uparrow$
SDEdit [22]	88.93	223.8	94.09 %	91.98%
Loopback [4]	104.6	132.9	54.28 %	85.32%
ILVR [5]	108.2	161.7	76.54 %	93.47%
Ours	94.40	134.2	N/A	N/A

Table 1. ***Quantitative Evaluations***. *(Left)* Method comparison w.r.t *faithfulness* $\mathcal{F}$ to the reference painting and *realism* $\mathcal{R}$ to the target domain. *(Right)* User-study results, showing % of inputs for which human subjects prefer our approach over prior works.

where $f(.)$ is the painting function. Thus an output image x is said to have high faithfulness with the given painting y if upon painting the final output x we get a painting $\tilde{y} = f(x)$ which is similar to the original target painting y.

Similarly, given a set of output data samples $\mathcal{S}(y, \tau_{text})$ conditioned on both painting y and text τ_{text}, and, $\mathcal{S}(\tau_{text})$ conditioned only on the text, the *realism* $\mathcal{R}$ is defined as,

$$\mathcal{R}(\mathcal{S}(y, \tau_{text})) = FID\left(\mathcal{S}(y, \tau_{text}), \mathcal{S}(\tau_{text})\right) \qquad (10)$$

where FID represents the Fisher inception distance [11].

Baselines. We compare our approach with prior works on guided image synthesis from stroke paintings with no paired data. In particular we show comparisons with, *1) SDEdit* [22] wherein the generative prior is introduced by first passing the painting y through the forward diffusion pass $y \rightarrow y_{t_0}$ [15, 35], and then performing reverse diffusion $y_{t_0} \rightarrow y_0$ to get the output image $x = y_0$[1]. *2) SDEdit + Loopback* [4] which reuses the last diffusion output to iteratively increase the realism of the final output. *3) ILVR*[2] [5]: which uses an iterative refinement approach for conditioning the output x of the diffusion model with a guidance image y. Unless otherwise specified, we use the GradOP+ algorithm (refer Alg. 2) when reporting evaluation results.

Qualitative Results. Results are shown in Fig. 4. We observe that both proposed approximate optimization methods (*i.e. GradOP* in row-1,2 and *GradOP+* in row-3,4) lead to output images which are both highly photorealistic as well as *faithful* with reference painting. In contrast, while SDEdit [22] shows high faithfulness to the input painting, the final outputs lack details and resemble more of a pictorial art rather than realistic photos. Iteratively reperforming the guided synthesis with the generated outputs (SDEdit + Loopback [4]) helps improve the realism of output images, however, we find that this has two main disadvantages. First, the iterative loop increases the effective time required for generating each data sample (*e.g.* four reverse

[1] We use standard hyperparameter value of $t_0 = 0.8$ in the main paper. Please refer supp. material for detailed comparisons for $t_0 \in [0, 1]$.

[2] Please note that the original ILVR [5] algorithm was proposed for iterative refinement with diffusion models in pixel space. We adapt the ILVR implementation for inference with latent diffusion models [28] for the purposes of this paper. Please refer supp. material for further details.

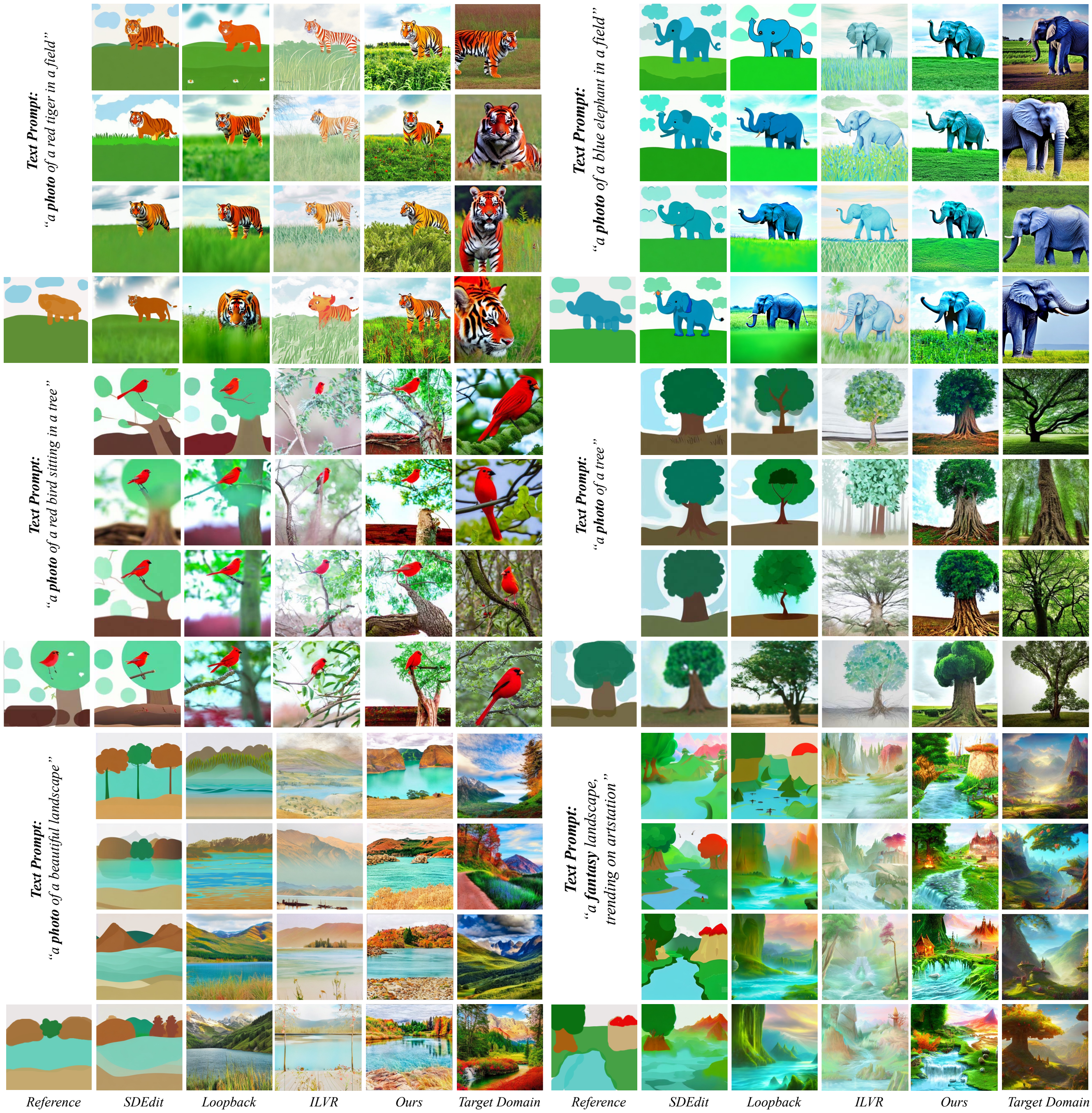


Figure 4. ***Qualitative comparisons.*** We compare the performance of our approach with prior works [4, 5, 22] based on their *faithfulness* to the provided reference, and the *realism* with respect to the target domain (generated by conditioning only on the text prompt). Please note that for our results, we show the *GradOP* (Alg. 1) and *GradOP+* (Alg. 2) outputs in row 1,2 and row-3,4 respectively.

sampling steps instead of just one). Second, we note that as the number of successive iterations increase the final outputs become less and less faithful to the original painting input. Finally, ILVR [5] leads to more realistic outputs, however, the final outputs are not fully faithful to the reference painting in terms of the overall color composition.

Quantitative Results. In addition to qualitative results we also quantitatively evaluate the final outputs on the *faithfulness* $\mathcal{F}(x, y)$ and *targeted-realism* $\mathcal{R}(.)$ metrics defined earlier. Additionally, similar to [22] we also perform a human user study wherein the *realism* and the overall satisfaction score (*faithfulness* + *realism*) are evaluated by human subjects (please refer supp. material for details). Results are shown in Tab. 1. We find that as expected, while SDEdit [22] leads to the best faithfulness with the target painting, it exhibits very poor performance in terms of the

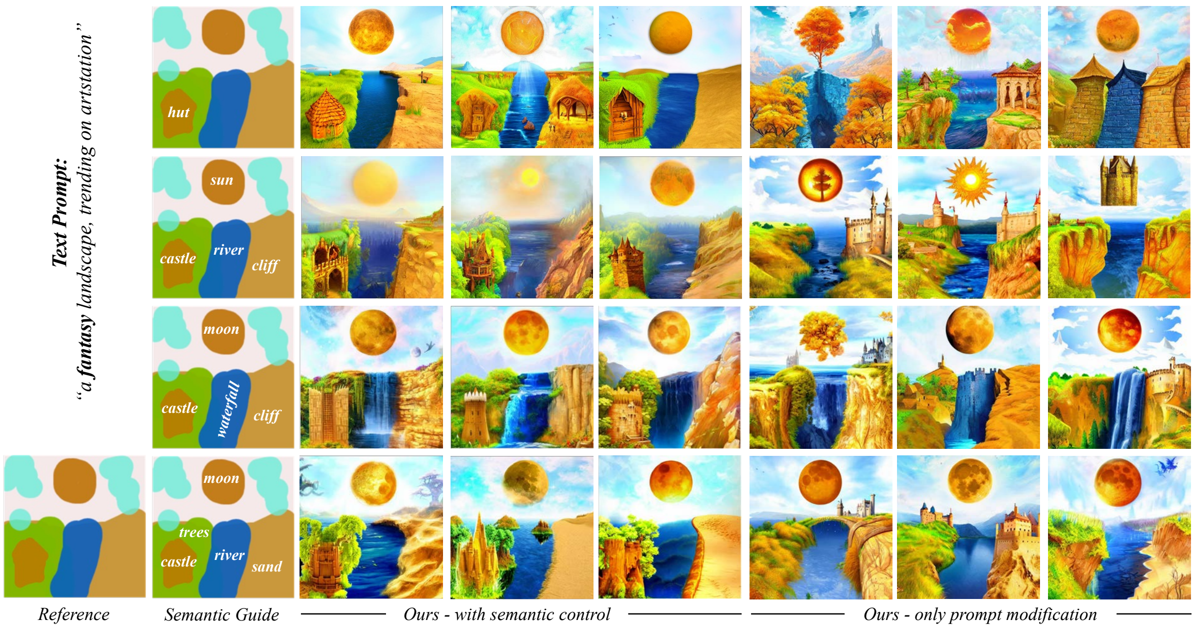


Figure 5. ***Controlling semantics of different painted regions.*** We compare image generation outputs (Col 3-5) using the cross-attention control approach from Sec. 3.3 with outputs (Col 6-8) generated by only modifying the input text prompt. Note that for each semantic guide (Col 2), the text prompt modification is performed by adding the corresponding semantic labels at the end of the text prompt. For instance, the modified text prompt for examples in row-1 would be "a fantasy landscape, trending on artstation showing a hut".

realism score. SDEdit with loopback [4] improves the realism score but the resulting images start loosing faithfulness with the given reference. In contrast, our approach leads to the best tradeoff between faithfulness to the target image and realism with respect to the target domain. These findings are also reflected in the user-study results wherein our method is preferred by $> 85.32\%$ of human subjects in terms of the overall satisfaction scores.

4.2. Controlling Semantics of Painted Regions

Results are shown in Fig. 5. We observe that in absence of semantic attention control, the model tries to infer the semantics of different painting regions in an implicit manner. For instance, the orange strokes in the sky region can be inferred as the sun, moon, or even as a yellow tree. Similarly, the brown strokes in the lower-left region (intended to draw a *hut* or small *castle*) are often inferred as muddy or rocky parts of the terrain. Moreover, such disparity continues even after modifying the input prompt to describe the intended semantic labels. For instance, in row-1 from Fig. 5, while changing the text prompt to include the text "hut" leads to the emergence of "hut" like structures, the inference is often done in a manner that is not intended by the user.

In contrast, by ensuring a high overlap between the intended painting regions and the cross-attention maps for the corresponding semantic labels (refer Sec. 3.3), we are able to generate outputs which follow the intended semantic guide in a much more accurate manner. For instance, the user is able to explicitly specify that the brown regions on the ground describes a hut (row 1) or castle (row 2-4). Similarly, the semantics of different regions can be controlled, *e.g*. the blue region is specified as a river or waterfall, the orange strokes in the sky is specified as moon or sun *etc*.

5. Analysis

5.1. Variation in Target Domain

In this section, we analyse the generalizability of the our approach across different target domains (*e.g*. children drawings, disney scenes) and compare the output performance with prior works. Results are shown in Fig. 6-a. We observe that our approach is able to adapt the final image outputs reliably across a range of target domains while still maintaining a high level of faithfulness with the target image. In contrast, SDEdit [22] generates outputs which lack details and thereby look very similar across a range of target domains. SDEdit + Loopback [4] addresses this problem to some extent, but it requires multiple reverse diffusion passes and the generated outputs tend to lose their faithfulness to the provided reference with each iteration.

5.2. Variation with Number of Gradient Steps

In this section, we analyse the variation in output performance as we change the number of gradient descent steps N_{grad} used to solve the unconstrained optimization problem in Sec. 3. Results are shown in Fig. 6-b. As expected, we find that for $N_{grad} = 0$, the generated outputs are sampled randomly from the subspace of outputs $(\mathcal{S}_{\tau_{text}})$ condi-

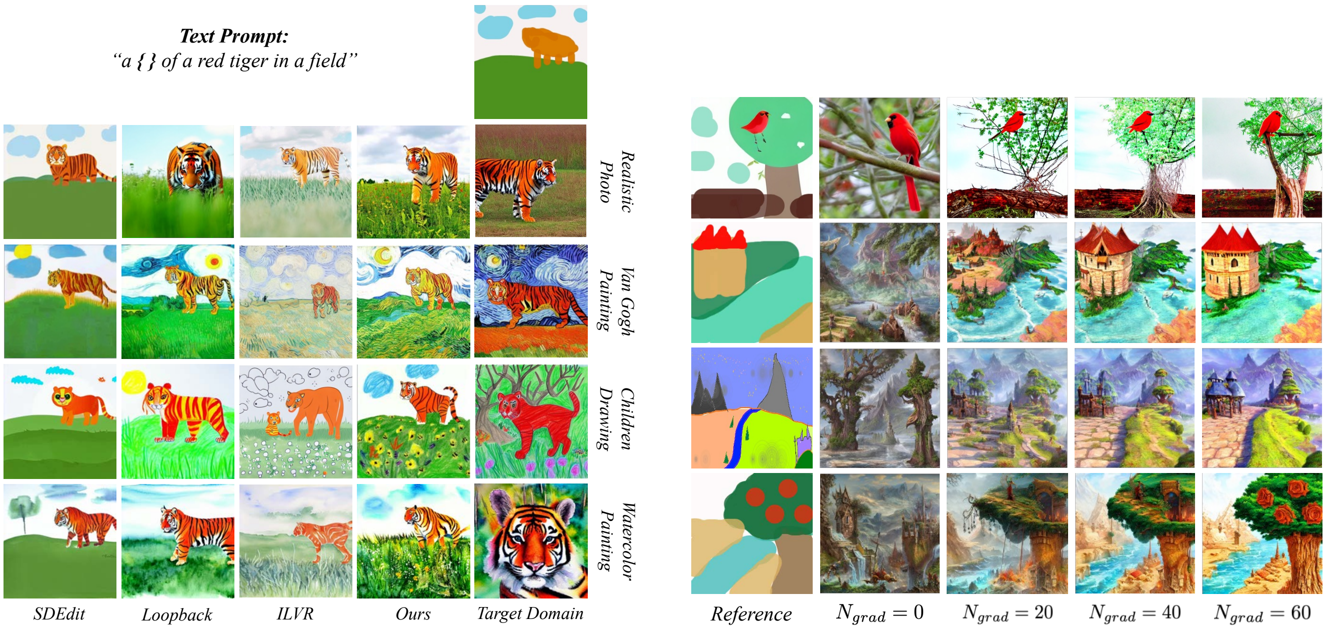


Figure 6. ***Method Analysis.*** Comparing guided image synthesis performance across *(left)* variation in target domain, and *(right)* variation in number of gradient descent steps N_{grad} used for performing the proposed optimization. Please zoom-in for best comparisons.

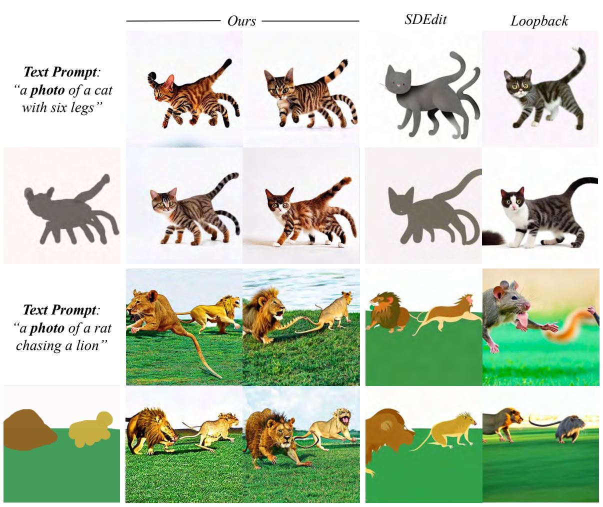


Figure 7. ***Out-of-distribution performance.*** Analysing *success* (top) and *failure* (bottom) cases for out-of-distribution prompts.

tioned only on the text. As the number of gradient-descent steps increase, the model converges to a subset of solutions within the target subsapce $\mathcal{S}_{\tau_{text}}$ which exhibit higher *faithfulness* with the provided reference. Please note that this behaviour is in contrast with SDEdit [22], wherein the increase in *faithfulness* to the reference is corresponded with a decrease in the *realism* of the generated outputs [22].

5.3. Out-of-Distribution Generalization

As shown in Sec. 4, 5, we find that the proposed approach allows for a high level of semantic control (both color composition and fine-grain semantics) over the output image attributes, while still maintaining the *realism* with respect to the target domain. Thus a natural question arises: *Can we use the proposed approach to generate realistic photos with out-of-distribution text prompts?*

As shown in Fig. 7, we observe that both success and failure cases exist for out-of-distribution prompts. For instance, while the model was able to generate "*realistic* photos of cats with six legs" (note that for the same inputs prior works either generate faithful but cartoon-like outputs, or, simply generate regular cats), it shows poor performance while generating "a photo of a rat chasing a lion".

6. Conclusions

In this paper, we present a novel framework for performing guided image synthesis synthesis with user scribbles, without the need for paired annotation data. We point that prior works in this direction [4, 5, 22], typically adopt an inversion-like approach which leads to outputs which lack details and are often simplistic representations of the target domain. To address this, we propose a novel formulation which models the guided synthesis output as the solution of a constrained optimization problem. While obtaining an exact solution to this optimization is infeasible, we propose two methods *GradOP* and *GradOP+* which try to obtain an approximate solution to the constrained optimization in a sample-efficient manner. Additionally, we show that by defining a cross-attention based correspondence between the input text tokens and user painting, it is possible to control semantics of different painted regions without the need for semantic segmentation based conditional training.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 2

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020. 2

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. 2

[4] AUTOMATIC1111. Stable-diffusion-webui. https://github.com/AUTOMATIC1111/stable-diffusion-webui, 2022. 2, 5, 6, 7, 8

[5] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 2, 4, 5, 6, 8

[6] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 2

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 2

[8] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022. 2

[9] 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. *arXiv preprint arXiv:2208.01618*, 2022. 2

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[12] Nisha Huang, Fan Tang, Weiming Dong, and Changsheng Xu. Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1085–1094, 2022. 2

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[14] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020. 2

[15] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 2, 5

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[17] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022. 2

[18] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 2

[19] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Ruifeng Deng, Xin Li, Errui Ding, and Hao Wang. Paint transformer: Feed forward neural painting with stroke prediction. *arXiv preprint arXiv:2108.03798*, 2021. 2

[20] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. *arXiv preprint arXiv:2112.05744*, 2021. 2

[21] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[22] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 2, 4, 5, 6, 7, 8

[23] 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. 1, 2

[24] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2

[25] 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. 5

[26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[27] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2, 5

[29] 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. *arXiv preprint arXiv:2208.12242*, 2022. 2

[30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 2

[31] Junyoung Seo, Gyuseong Lee, Seokju Cho, Jiyoung Lee, and Seungryong Kim. Midms: Matching interleaved diffusion models for exemplar-based image translation. *arXiv preprint arXiv:2209.11047*, 2022. 2

[32] Jaskirat Singh, Cameron Smith, Jose Echevarria, and Liang Zheng. Intelli-paint: Towards developing human-like painting agents. In *European conference on computer vision*. Springer, 2022. 2

[33] Jaskirat Singh and Liang Zheng. Combining semantic guidance and deep reinforcement learning for generating human level paintings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[34] Jaskirat Singh, Liang Zheng, Cameron Smith, and Jose Echevarria. Paint2pix: Interactive painting based progressive image synthesis and editing. In *European conference on computer vision*. Springer, 2022. 2

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3, 5

[36] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781*, 2020. 2

[37] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *arXiv preprint arXiv:2102.02766*, 2021. 2

[38] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022. 5

[39] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 2

[40] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 1, 2

[41] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022. 2

[42] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pages 597–613. Springer, 2016. 2

[43] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 2

[44] Zhengxia Zou, Tianyang Shi, Shuang Qiu, Yi Yuan, and Zhenwei Shi. Stylized neural painting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15689–15698, 2021. 2