

LDEdit: Towards Generalized Text Guided Image Manipulation via Latent Diffusion Models

Paramanand Chandramouli
paramanand.chandramouli@uni-siegen.de

Department of Computer Science
University of Siegen
Germany

Kanchana Vaishnavi Gandikota
kanchana.gandikota@uni-siegen.de

Abstract

Research in vision-language models has seen rapid developments off-late, enabling natural language-based interfaces for image generation and manipulation. Many existing text guided manipulation techniques are restricted to specific classes of images, and often require fine-tuning to transfer to a different style or domain. Nevertheless, generic image manipulation using a single model with flexible text inputs is highly desirable. Recent work addresses this task by guiding generative models trained on the generic image datasets using pretrained vision-language encoders. While promising, this approach requires expensive optimization for each input. In this work, we propose an optimization-free method for the task of generic image manipulation from text prompts. Our approach exploits recent Latent Diffusion Models (LDM) for text to image generation to achieve zero-shot text guided manipulation. We employ a deterministic forward diffusion in a lower dimensional latent space, and the desired manipulation is achieved by simply providing the target text to condition the reverse diffusion process. We refer to our approach as LDEdit. We demonstrate the applicability of our method on semantic image manipulation and artistic style transfer. Our method can accomplish image manipulation on diverse domains and enables editing multiple attributes in a straightforward fashion. Extensive experiments demonstrate the benefit of our approach over competing baselines.

1 Introduction

Using natural language descriptions is an intuitive and easy way for humans to communicate visual concepts. Hence, a tool which can automatically manipulate images using textual descriptions can greatly ease editing. This requires a careful control to modify only the relevant semantic attributes and styles while preserving the desired content representations. However, accomplishing this is highly challenging, especially when manipulating open-domain images using arbitrary text prompts. As a result, many existing works allow manipulations which are restricted to a specific image classes [28, 40, 56, 51] or a specific manipulation task [8, 12, 22]. Further, some of these methods require fine-tuned models [28, 40] for specific text prompts, further limiting their utility for flexible open domain image manipulation. In contrast to these techniques, the works [12, 22] handle general image manipulation from text prompts. While

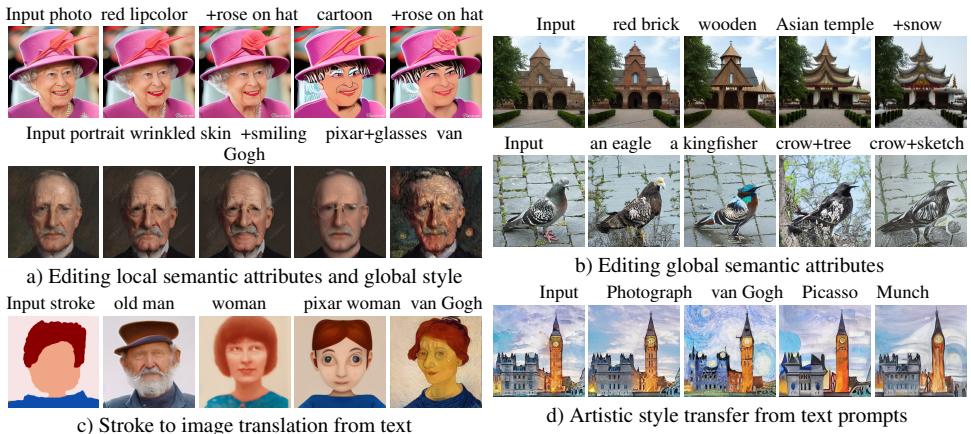


Figure 1: LDEDIT can edit local and global semantic attributes and also perform artistic style transfer on real-world images using a single model.

[\textcolor{red}{\texttt{L}}] focuses on semantically simple transformations, [\textcolor{red}{\texttt{G}}] allows more general text-to-image generation as well as manipulation using an expensive latent space optimization.

In this work, we attempt to develop a fast and flexible approach to open domain image manipulation using arbitrary text prompts. Our goal is to accomplish a wide range of manipulations from text prompts ranging from simple change in colour of an object, to modification of multiple semantic attributes of image, and artistic styles, all with a single model. Our work is inspired by the recent dramatic developments in realistic image generation with language guidance [\textcolor{red}{\texttt{D}}, \textcolor{red}{\texttt{E}}, \textcolor{red}{\texttt{F}}, \textcolor{red}{\texttt{G}}]. In particular, we leverage the recently proposed Latent Diffusion Model (LDM) [\textcolor{red}{\texttt{B}}] which performs diffusion in a smaller dimensional latent space of trained convolutional auto-encoders, to provide higher inference speed and computational efficiency. Further, we utilize the idea of non-Markovian diffusion proposed in Denoising Diffusion Implicit Models (DDIM) [\textcolor{red}{\texttt{H}}] which can enable faster inference and high fidelity sample reconstruction. Our key idea is the use of a shared latent representation as a link between the source image and the desired target. To this end, we employ a deterministic DDIM sampling in the forward diffusion in the latent space of LDM. We use the same latent code along with the target text prompt to condition the reverse diffusion process, effectively achieving desired transformation in the input image, while automatically maintaining consistency with the original content representation. Using this technique, we can accomplish a variety of image manipulation tasks using the pretrained LDM, in a zero-shot fashion without further optimization or fine-tuning. Further, by introducing controlled stochasticity, we can trade-off diversity for fidelity with original image. This is especially useful when the desired target is very different from the original input.

Fig 1 illustrates the diverse image editing tasks that can be accomplished by our LDEDIT using only text prompts. We can modify objects in the image while largely preserving the original pose or structure, see Fig. 1 b). LDEDIT can accomplish simultaneous global style manipulation as well as fine-grained (multiple) attribute changes such as change in expression, wrinkles, makeup while preserving identity in human faces, see Fig. 1 a). Further, without requiring an input mask, simple local edits such as adding a flower on a woman’s hat, or eye glasses are achieved through text alone. Our approach can operate on diverse types of input images such as natural photographs, paintings, sketches, and strokes. By providing an

Method	Image input	Text input	Semantic (Global)	Artistic Style	local edits	Comments
DiffusionCLIP [2]	Class-Specific	Predefined	✓	✓	✗	Separately fine-tuned models for each task
StyleCLIP [3]	Class-specific	Arbitrary	✓	✗	✗	Includes versions with and without optimization
GLIDE [4]	Open domain	Arbitrary	✗	✗	✓	Trained model for inpainting with mask input
CLIPStyler [5]	Open domain	Arbitrary	✗	✓	✗	Test-time optimization w/o pretrained generator
VQGAN+CLIP [6]	Open domain	Arbitrary	✓	✓	Limited	Optimization with pretrained generator
LDEdit(Ours)	Open-domain	Arbitrary	✓	✓	✓	A pretrained LDM is used

Table 1: Comparison of recent state of the methods for text guided image manipulation.

intuitive target text prompt " a photograph of a woman" or a "pixar animation of a woman", our method can translate from stroke to a semantically consistent image in the corresponding domain, see Fig. 1 c). We can observe realistic details are hallucinated while transferring to the domain of natural photos, for example, wrinkles in the picture of old man, in Fig. 1 c), or details in the clock Fig. 1 d). Further, artistic style transfer is also achieved via simple text prompts, such as "a Picasso style painting". It can be seen that our approach can accomplish manipulations that are semantically and stylistically consistent with the given target text prompt, while remaining faithful to original content.

By offering significant advantages in flexibility, faster run-times and capability to generate diverse samples in parallel, LDEdit can facilitate efficient user-guided editing. Our experimental results demonstrate that LDEdit can accomplish diverse manipulation tasks, in addition to achieving performance close to recent state of the art baselines.

2 Related Work

Image Generative Models Ever since the seminal works of VAEs [41] and GANs [31], image generative models have achieved significant improvements, and modern generative models can generate highly photo-realistic images [2, 20, 25, 28, 29, 30, 20]. While GANs [31] achieve high quality generation, they are difficult to train and are prone to mode collapse. Likelihood-based models, [42, 60] on the other hand, have a stable training and capture more diversity. Score based [21, 22] or denoising diffusion [23, 25] models are a new class of likelihood-based models built from a hierarchy of denoising auto-encoders [28]. These models have recently demonstrated generative capabilities surpassing GANs [24, 23]. Yet, high quality diffusion models are computationally expensive to train, and have slower inference times than GANs, due to expensive Markovian sampling and iterative network evaluations required for diffusion. These problems can be alleviated by accelerated stochastic sampling techniques or by performing diffusion in a smaller latent space [63, 76]. Employing deterministic diffusion process [71] can also speed up inference, in addition to enabling high fidelity sample reconstruction, which can be exploited for image recovery and manipulation.

Image Manipulation As images can be manipulated in various ways, (*e.g.* artistic style, image translation, semantic manipulation, local edits), a variety of methods exist. Approaches for image translation include CNN based optimization using style and content images [30], conditional GANs trained on pair of domains [1, 26, 27, 20], GANs for multi-domain translation [12, 13] and more recently, conditional diffusion models [62, 65]. An alternate approach [2, 29] is to manipulate images in the latent space of pretrained GANs. StyleGANs [28, 29] are a popular choice for such latent space editing due to their disentanglement properties in the latent space [1, 16, 23, 28, 20, 23]. This is achieved through optimization or by using encoders for GAN inversion [3, 62, 23]. However, GAN inversion may not yield faithful reconstruction [3]. Improving StyleGAN inversion for editing is an active area of research [3, 1, 22, 25, 29].

In contrast to GANs, diffusion models can readily be leveraged for inpainting [43] and stroke guided image editing [50] and even unpaired image translation [73].

Text Guided Generation and Manipulation: Earlier works employed RNNs [49] and GANs [43, 61, 82, 84, 85, 83, 86, 91, 92] for text guided image synthesis, and manipulation [23, 44, 52]. Nevertheless, these works are often restricted to class specific image generation and are trained on smaller datasets. In the recent past, there is a rapid surge in vision-language models, with the developments in cross-modal contrastive learning [37, 57] and powerful text-to-image generative models [24, 58, 65, 66]. These models are trained on massive datasets to learn joint image-text distributions. Some of these models [21, 27, 58] use autoregressive(AR) transformers for generation, while some others [54, 59, 66] employ diffusion based models for the generation task. However, training these models for high quality generation requires massive computational resources. To address this, some recent works [10, 24, 43, 45, 63, 74] instead perform the diffusion in a lower dimensional latent space resulting in faster training and inference. In our work, we exploit Latent Diffusion Models (LDM) [63] as they offer good reconstruction quality, latency, and perform diffusion in a continuous latent space.

CLIP [57] is a cross modal encoder which provides a similarity score between an image and a caption. Several recent approaches to text guided image synthesis [10, 18, 19, 29, 46, 47, 51, 52] steer pretrained generative models [20, 21, 25] towards a user provided text prompts using CLIP. This approach of CLIP controlled latent space navigation is directly applicable for image manipulation [19], mask guided local editing [8, 9], semantic manipulation of class-specific images [2, 56, 83] via StyleGAN inversion [8]. CLIP has also been applied to fine-tune output domain and style [28, 40] of class-specific image generators. While these approaches are promising, optimization in latent space for each text-prompt is expensive and time-consuming. On the other hand, the fine-tuned models are fast, but restricted to the specific fine-tuned tasks. Further, class-specific generators are not suited for manipulation of open domain images. Instead of using pretrained generative models, some recent works employ test-time optimization for each image and target text, using CLIP, for tasks such as local object appearance [2], global texture-style manipulation [42], rendering drawings [13, 26], however such optimization is task specific, and is expensive requiring many augmentations. Tab. 1 provides an overview comparing the pros and cons of recent methods for text guided manipulation. As we can see, our approach and VQGAN+CLIP [19] can accomplish flexible manipulation tasks. Additionally, our approach allows fast manipulations.

3 Preliminaries

Diffusion Models: Denoising diffusion probabilistic models (DDPM) [52] are characterized by two diffusion processes: i) a forward process to gradually corrupt data samples into a tractable distribution e.g. Gaussian distribution, ii) a learned iterative denoising process to convert Gaussian noise to samples from data distribution. The forward diffusion involves progressively noising a clean image x_0 in T time-steps with transitions $q(x_t | x_{t-1}) := \mathcal{N}(\sqrt{1 - \beta_t}x_{t-1}, \beta_t \mathbf{I})$, where $\{\beta_t\}_{t=0}^T$ is the noise variance schedule. The evolution of x_t can be expressed as

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{(1 - \alpha_t)}\zeta, \quad \text{where } \zeta \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \text{ and } \alpha_t := \prod_{s=1}^t (1 - \beta_s). \quad (1)$$

The generative process progressively denoises x_T to x_0 also via a Gaussian transition which is approximated by learned network ϵ_θ .

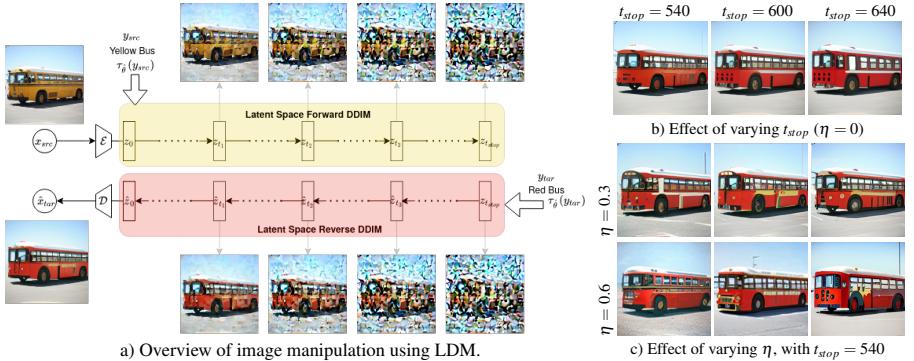


Figure 2: a) Overview of LDEdit, illustrating forward and reverse diffusion in latent space of autoencoder. b) and c) illustrate the effects of varying time steps t_{stop} and stochasticity hyperparameter η respectively

The reverse diffusion process is expressed as:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} \boldsymbol{\epsilon}_\theta(x_t, t) \right) + \sigma_t \xi, \quad \text{where } \xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (2)$$

Denoising Diffusion Implicit Models(DDIM) [20] employ a different non-Markovian forward process with the same forward marginals as DDPM:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1-\alpha_t} \boldsymbol{\epsilon}_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \boldsymbol{\epsilon}_\theta(x_t, t) + \sigma_t^2 \xi, \quad (3)$$

where $\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\alpha_0 := 1$, by definition. Varying σ leads to different generative processes with the same model $\boldsymbol{\epsilon}_\theta$. When σ_t is set to 0, the DDIM sampling becomes fully deterministic, enabling fast inversion of the noised latent variable to the original images (or to x_0 in our case) [20, 21]. In this case, the deterministic forward DDIM process expressed as:

$$x_{t+1} = \sqrt{\alpha_{t+1}} \left(\frac{x_t - \sqrt{1-\alpha_t} \boldsymbol{\epsilon}_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) + \sqrt{1-\alpha_{t+1}} \boldsymbol{\epsilon}_\theta(x_t, t) \quad (4)$$

and the deterministic reverse DDIM process is expressed as:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1-\alpha_t} \boldsymbol{\epsilon}_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) + \sqrt{1-\alpha_{t-1}} \boldsymbol{\epsilon}_\theta(x_t, t) \quad (5)$$

For different subsequences τ in $[1, \dots, T]$ [20] consider σ of the form:

$$\sigma_\tau(\eta) = \eta \sqrt{(1-\alpha_{\tau_{i-1}})/(1-\alpha_{\tau_i})} \sqrt{1-\alpha_{\tau_i}/\alpha_{\tau_{i-1}}}, \quad (6)$$

where the hyperparameter $\eta \in \mathbb{R}_{\geq 0}$ controls the degree of stochasticity, with $\eta = 1$ leading to original DDPM generative process and $\eta = 0$ leading to DDIM.

Latent Diffusion Models: The main idea of LDMs is to perform diffusion in the latent space of an autoencoder to improve speed and computational efficiency. Given an image $x_{src} \in \mathbb{R}^{H \times W \times C}$, the encoder \mathcal{E} maps x_{src} into a down-sampled latent code $z_0 = \mathcal{E}(x_{src})$, and the decoder \mathcal{D} is trained to recover the image from this latent. This encoding results in a lossy compression, i.e. $\|\mathcal{D}(\mathcal{E}(x_{src})) - x_{src}\|$ is finite, which is a trade-off for computational efficiency. Following encoding into latent space, diffusion process can happen via DDPM or

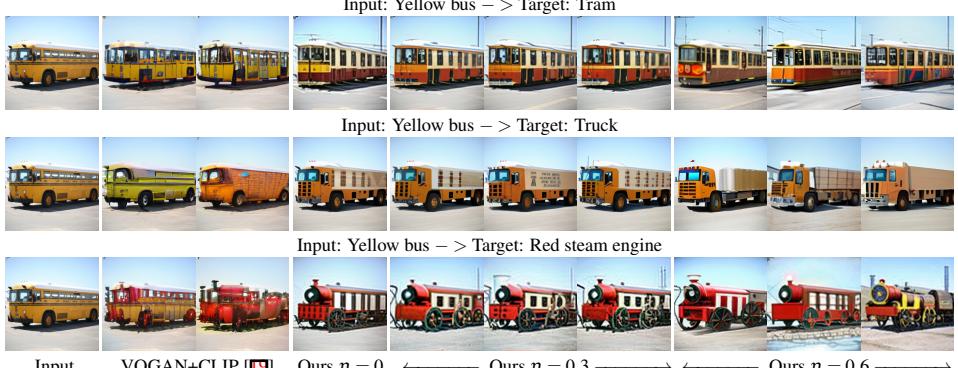


Figure 3: Comparison with VQGAN+CLIP [19]: Manipulation results of yellow bus according to target texts ‘a tram’, ‘a truck’ and ‘a red steam engine’.

DDIM (1)–(5), but in z_t for $t \in [1, T]$ instead of x_t . The diffusion process can additionally be conditioned on user inputs such as text prompts $\boldsymbol{\epsilon}_\theta(z_t, t, \tau_{\tilde{\theta}}(y))$. Here, the text-prompts y are tokenized using transformers $\tau_{\tilde{\theta}}$ [7] for conditioning the diffusion process.

4 Text Driven Manipulation with LDEdit

In this section, we show how LDMs trained for text-to-image generation can be adapted for image manipulation. Our main idea is to use a common shared latent representation between the source image and the desired target, which is made possible by a deterministic diffusion process. The source image x_{src} is mapped to a latent code z_0 by the encoder \mathcal{E} , and forward diffusion is performed until the time step $t_{stop} < T$ using DDIM sampling, conditioned on the source text prompt y_{src} as:

$$z_{t+1} = \sqrt{\alpha_{t+1}} \left(\frac{z_t - \sqrt{1 - \alpha_t} \boldsymbol{\epsilon}_\theta(z_t, t, \tau_{\tilde{\theta}}(y_{src}))}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t+1}} \boldsymbol{\epsilon}_\theta(z_t, t, \tau_{\tilde{\theta}}(y_{src}))$$

The reverse diffusion conditioned on the target text prompt y_{tar} starts from the same noised latent code $z_{t_{stop}}$ to arrive at \hat{z}_0 :

$$z_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{z_t - \sqrt{1 - \alpha_t} \boldsymbol{\epsilon}_\theta(z_t, t, \tau_{\tilde{\theta}}(y_{tar}))}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1}} \boldsymbol{\epsilon}_\theta(y_t, t, \tau_{\tilde{\theta}}(y_{tar})) \quad (7)$$

Due to deterministic sampling, a near cycle-consistency is automatically maintained between source and target images [23]. Fig. 2 a) provides an overview of our approach, with an example where a source image with y_{src} ‘a yellow bus’, is transformed according to the y_{tar} ‘a red bus’ in a straightforward way. The visualized results obtained by decoding latents sampled in $[1, t_{stop}]$ during the forward and reverse diffusion process demonstrate the gradual transformation in the reverse process. Additionally, we can also introduce controlled stochasticity by varying η (6), which can produce diverse outputs as seen in Fig. 2 c), with magnitude of η controlling consistency with the original image. Further, Fig. 2 b) shows that changing the number of DDIM steps can also lead to some variance in our results. In the following section, we demonstrate that this technique can accomplish a variety of image manipulation tasks using the pretrained LDM, in a zero-shot fashion without further optimization or fine-tuning.

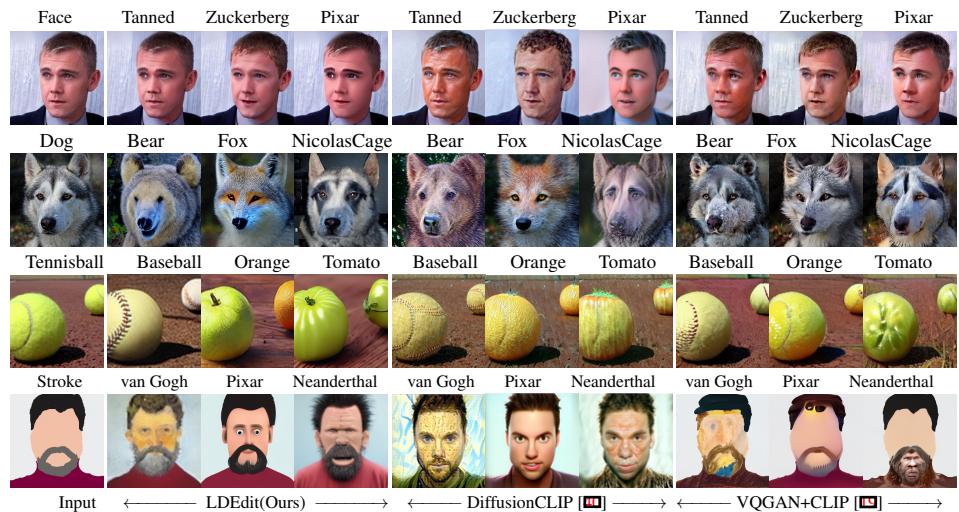


Figure 4: Visual comparison of image manipulation task with DiffusionCLIP [40] and VQGAN+CLIP [19]. Our LDEdit can successfully transform input image into target classes while retaining the original pose.

5 Experiments

We perform all our experiments with different image manipulation tasks using the text-to-image LDM with a downsampling factor of 8 pretrained using the openly available LAION dataset [57] containing open-domain image-text pairs. We do not fine-tune this model for any task. We set $t_{stop} \in [300, 640]$ out of the total 1000 steps and use fewer (20-80) steps between $[1, t_{stop}]$ in the deterministic forward and reverse diffusion. We perform experiments on both class-specific and open-domain images and compare with VQ+CLIP [19] which is versatile to handle general manipulation tasks. In addition, we also compare with class-specific approaches [56, 81] and fine-tuned models [28, 40] on the domain-specific tasks. Comparisons with the baseline-methods and run-time comparisons are performed with images of dimension 256×256 .

We first demonstrate our method on the task of manipulating an image of a yellow bus according to the target prompts: ‘a tram’, ‘a truck’ and ‘a red steam engine’. Fig. 3 illustrates the results of this manipulation. The results indicate that LDEdit is able to manipulate the input according to the target texts even with a simple DDIM forward and reverse process with $\eta = 0$. Further, by increasing η , our method is able to generate an assortment of diverse samples that are consistent with the pose of the yellow bus in the input image. The diversity increases as the parameter η is increased. We also illustrate the results obtained by VQGAN+CLIP [19] on this task using two sets of hyper-parameters for comparison. While [19] can successfully transform the input image to that of ‘a tram’, we were unable to obtain satisfactory results for the other two tasks, despite manual hyper-parameter tuning.

We further test our approach on manipulating images from diverse classes using test images from [40]. We compare our performance with the generic approach of VQGAN+CLIP [19] and DiffusionCLIP [40], a state of the art method using class-specific models fine-tuned for the specific target texts. Fig. 4 illustrates the results of this experiment. As DiffusionCLIP uses specific fine-tuned models on these tasks, it can effortlessly accomplish the desired

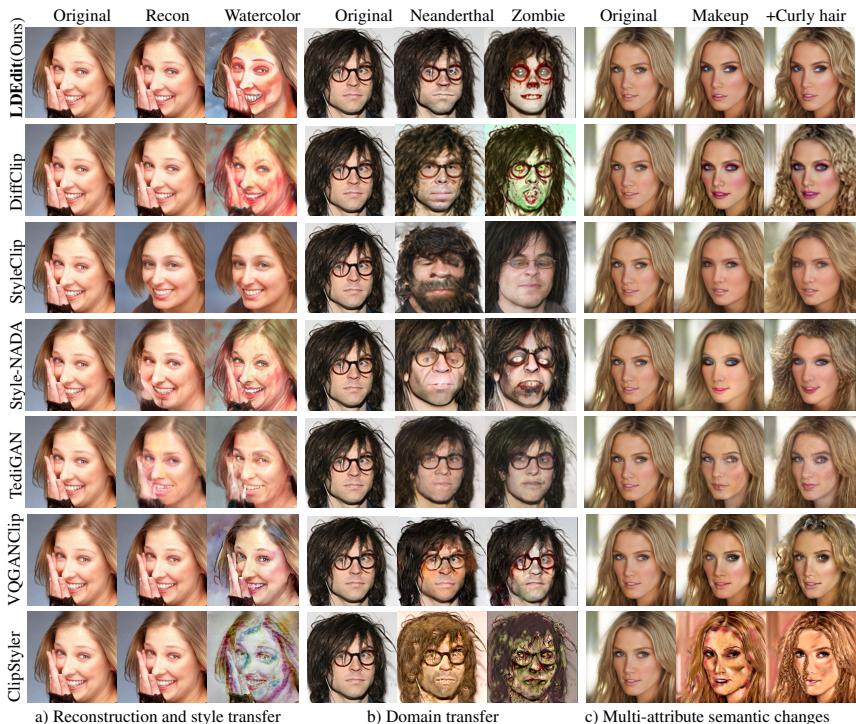


Figure 5: Comparison with recent baselines: DiffusionCLIP [40] StyleCLIP [56], StyleGAN-NADA [28], TEDIGAN [31], CLIPStyler [42], VQGAN+CLIP [19].



Figure 6: Simultaneous editing of multiple attributes and objects of an image. Shown from left to right are (i) input (ii) girl+watermelon (iii) woman+corgi (iv) paint+cat+old woman (v) paint + boy+ big egg (vi) paint + man + rabbit (vii) paint + man + dog (viii) man+cat

manipulations. On the other hand, VQGAN+CLIP struggles to achieve desired changes when the target is highly different from the input. Despite not being fine-tuned for the specific tasks, our LDEdit can accomplish the manipulations quite well. The task of manipulating a stroke image according to the target prompts is particularly challenging, as the input image lacks details. Handling such manipulation requires introducing stochasticity in the forward process, without which it is not possible to produce the desired edits.

We further perform multiple manipulation tasks on face images, including semantic (multi)-attribute manipulation, style transfer, domain manipulation and compare with the recent state-of-the-art methods which are trained for face manipulation [28, 40, 56, 31]. The StyleGAN based methods [28, 56, 31] employ the same encoders for GAN inversion as per the original setting in their work. Further, we include comparison with CLIP-Styler [42] a CLIP guided texture manipulation approach, and VQGAN+CLIP [19] which can perform flexible image manipulation. Fig. 5 illustrates our results. While StyleGAN inversion

based approaches [28, 26, 31] can manipulate semantic attributes see Fig.5 c), they struggle to reconstruct face images in atypical poses, see Fig.5 a). Unexpected details present in the original image such as hand on the face are completely removed or distorted in the reconstructions. Since such atypical faces are hardly encountered during training, StyleGAN inversion results in a high representation error. Similarly, it is hard to transfer to a different style *e.g.* a watercolour painting, or domain *e.g.* zombie using StyleGAN latent space search alone Fig.5 a) and b). StyleGAN-NADA instead enable these manipulations using domain-specific fine-tuning. On the other hand, ClipStyler [20] can only accomplish global texture manipulations, and the result may drift away from the original colour palette. Among the compared methods, LDEdit, DiffusionCLIP [40] and VQGAN+CLIP[19] accomplish the different manipulation tasks in addition to achieving good reconstructions, preserving identity better than GAN inversion based methods. Interestingly, though VQGAN+CLIP and LDEdit are trained on generic images, these methods are still able to perform on par with state of the art fine-tuned DiffusionCLIP [40] on these tasks.

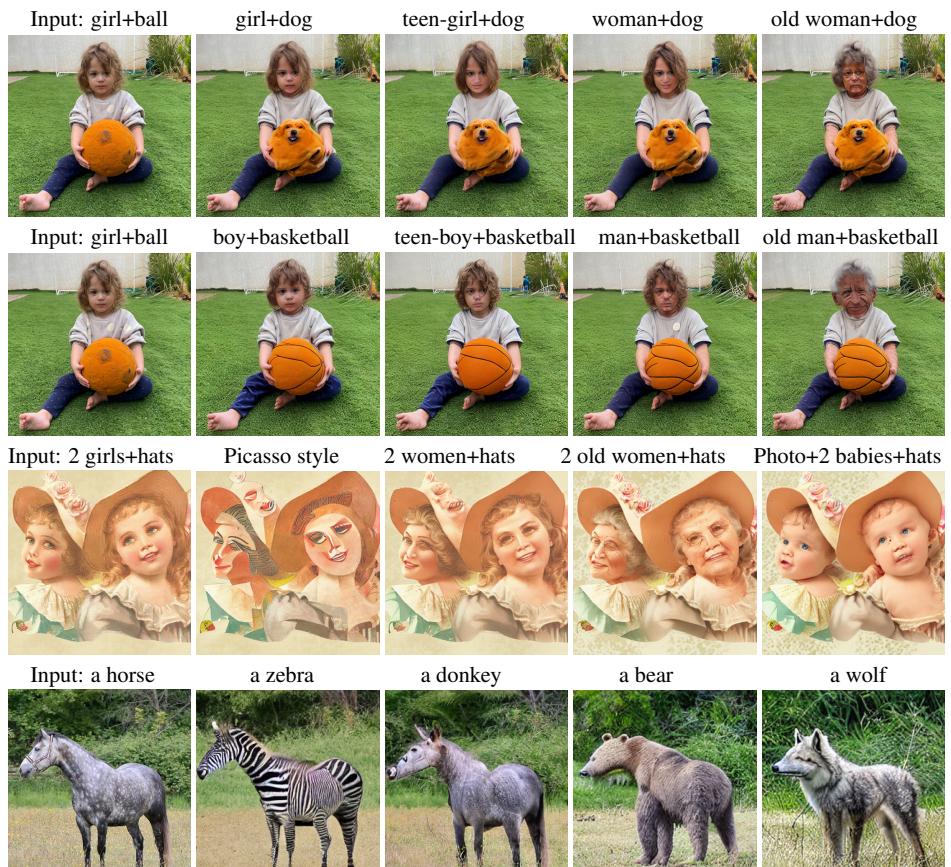


Figure 7: Results of image manipulation using LDEdit

It is also possible to achieve further challenging manipulations involving simultaneous changes in multiple attributes, local manipulations and artistic style changes as seen in Fig. 6. While the LDM model is trained on generation of images of dimension 256×256 , due to

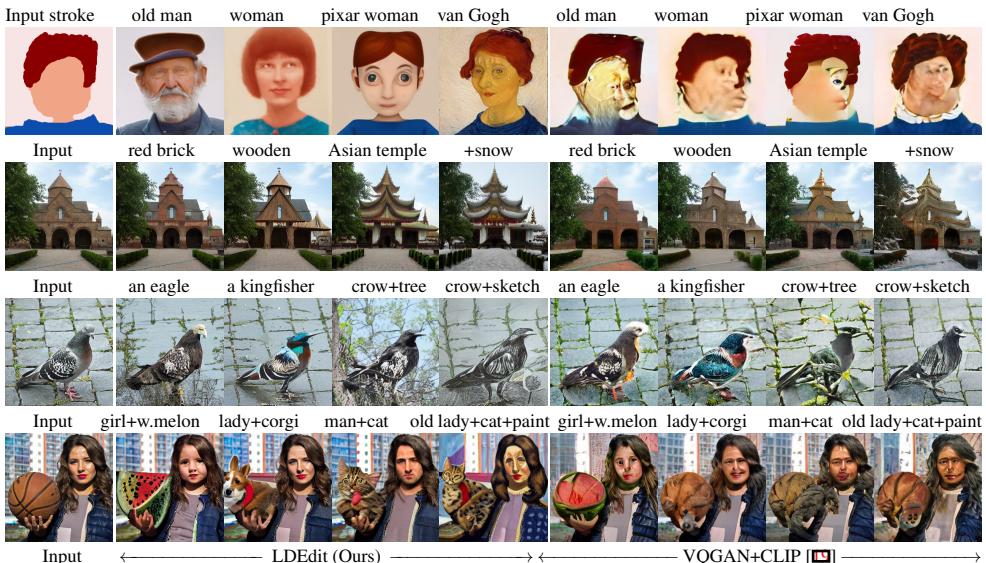


Figure 8: Comparison of LDEdit with VQGAN+CLIP [19]. Best results out of 4 samples are shown for LDEdit and the best result out of 8 samples are shown for VQGAN+CLIP.

fully convolutional nature of the autoencoder, our method can be applied on images of higher resolution using the same model. Fig. 7 shows further example results of image manipulation using LDEdit, with image resolution 512×512 . It is seen that our method can achieve varied transformations in a straightforward way. The first two rows show simultaneous manipulation of the girl and the ball. The third row shows style transfer to a painting or a photo and semantic manipulating the age of two girls. Interestingly, LDEdit can effect such transformations with a little or no stochasticity, such that the background remains largely unaffected. The final row shows manipulating a horse to other species, *e.g.* a zebra, a donkey, a bear, and a wolf. These transformations required a higher η of 0.3 for zebra and donkey, and η of 0.8 for bear and wolf. However, higher values of η result in more changes in the background.

Comparisons with VQGAN+CLIP [19] We provide more visual comparisons with VQGAN+CLIP for general text driven image manipulation in Fig. 8. While VQGAN+CLIP can successfully effect changes in the input image of a building as per the target text prompts, its performance suffers in more difficult manipulations such as translating from an input stroke, or performing simultaneous local manipulations. In contrast, our LDEdit is able to perform these desired manipulations.

Failure Cases In some cases, our method may fail to produce desired manipulations as seen in Fig. 9. With an input text prompt of ‘a deer with antlers’, we obtain manipulated images where the antlers are misplaced. In other cases, we obtain features of target objects additionally in undesired locations, such as a baby face on the girl’s hand, or a cat face in the hair and in the background picture frame. These undesired effects can be avoided by using a mask, which can aid in localization of edits.

Editing with Masks Our method can be modified to include a user-specified mask which specifies the regions where significant changes are needed. Similar mask-guided editing has also been shown in [8, 24]. The user-specified mask is also down-sampled such that it has the same spatial extent as the latent code. Let $z_{t_{stop}}$ be the latent code after forward diffusion,

the desired localized edit can be obtained by performing the reverse diffusion process on multiple copies of $z_{t_{stop}}$, by changing the target text for the respective masked regions. For seamless blending of the masked and unmasked regions, the latent code corresponding to the two regions are combined at each diffusion step. This even allows us to specify different levels of stochasticity for the different regions. Fig. 10 shows the result of such mask masked editing. We can see that our approach successfully results in a seamless local editing, without requiring expensive optimization.

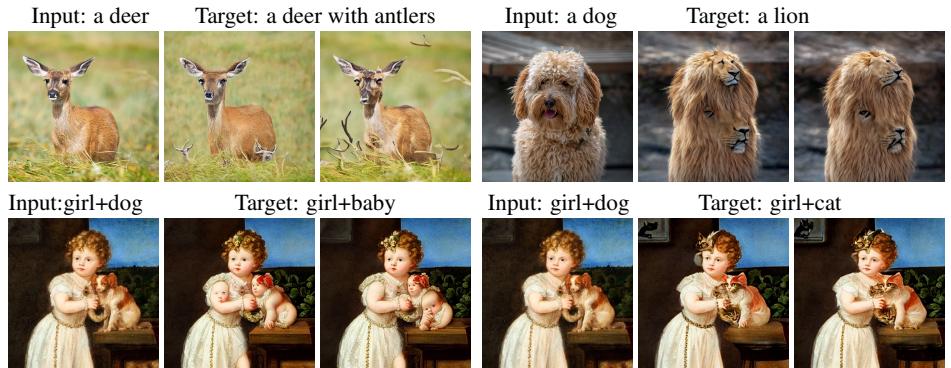


Figure 9: Failure cases of image manipulation using LDEdit



Figure 10: Masked image manipulation using LDEdit

Effect of Stochasticity In our approach, we proposed to perform a deterministic DDIM sampling, to ensure that a consistency is maintained with the original image. However, when the input image lacks details, such as a stroke image, doing a deterministic forward produces a latent code which lacks any details, see Fig. 11 a). On the other hand, introduction of stochasticity through η can aid in hallucinating details not present in the original image, Fig. 11 b). With $\eta = 1$, DDIM becomes equivalent to DDPM sampling, which results in more diverse samples. Note that our method may sometimes result in images with text like artifacts, as seen in Fig. 11 c). More example image manipulation of LDEdit by varying η are shown in Fig. 12. As the value of η increases, the diversity of samples improves. However, there are more perceptible changes in background, see rows 1 and 2 of Fig. 12.

User Study We conduct user studies to compare user preference of image manipulation results of our method with VQGAN+CLIP [49] and DiffusionCLIP [40]. Users participated in two surveys, where they were provided with source image, target text description and the results obtained with LDEdit and base-line method (VQGAN+CLIP or DiffusionCLIP) in a random order, and voted their preferred image manipulation using a survey platform. We obtained a total of 1120 votes from 32 participants for comparing LDEdit with VQGAN+CLIP and

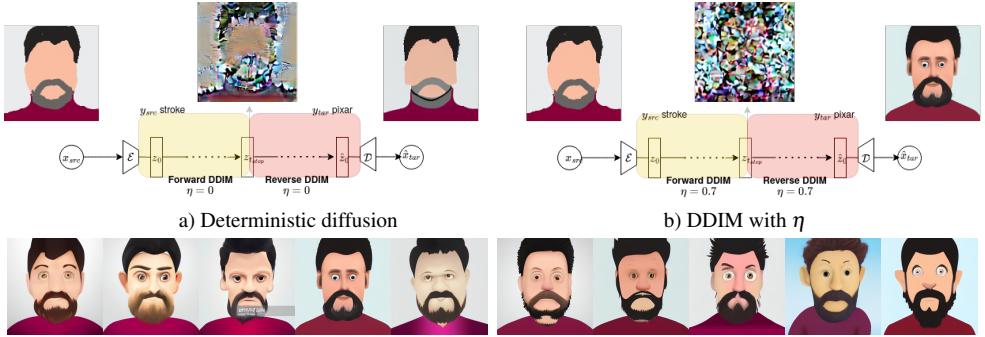


Figure 11: Effect of η in diffusion process. Purely deterministic DDIM process cannot achieve desired target when the original input lacks details.

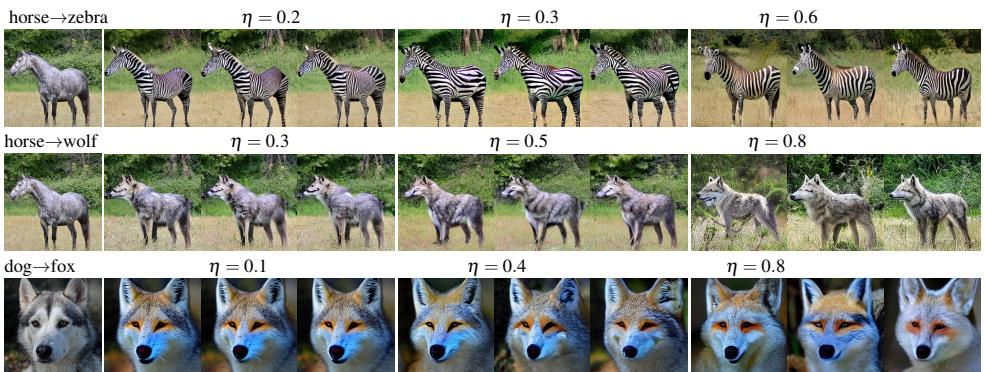


Figure 12: Sample results for different η using LDEdit. As the value of η increases, the diversity of samples increases.

950 votes from 38 participants for comparing LDEdit with DiffusionCLIP. For comparison with both the baselines, we included a combination of face images and general images (on manipulations demonstrated in DiffusionCLIP [40] paper). On faces, manipulated attributes include makeup, tanned, curly hair, changing gender, domain change to zombie, neanderthal. We also include an example of translating stroke image to pixar, neanderthal and van Gogh painting. On general image manipulation, we include manipulating an input building, bus, dog and a tennis ball. Additionally, for comparison with VQGAN+CLIP, we include examples of manipulating an image of a bird and multiple local object manipulations. In human evaluation, the results of LDEdit were preferred 83.87% of the time in the survey comparing LDEdit with VQGAN+CLIP, whereas user preference for LDEdit is 49.15% in the survey comparing LDEdit with DiffusionCLIP.

Run-time Tab. 2 provides a comparison of GPU memory requirements and run-times of different text based image manipulation methods. The experiments were conducted on a computer with AMD Ryzen 9 3950X 16-Core Processor and NVIDIA GeForce RTX 3090 with 24GB GPU memory. The run-times are highest for VQGAN+CLIP [49] (in the order of minutes), which requires an expensive optimization. Further, VQGAN+CLIP requires different number of iterations to achieve the desired edit depending on the target prompt, leading to variable run-times. The run-times of both DiffusionCLIP [40] and our proposed

LDEdit are significantly lower, with LDEdit having smaller run-times due to diffusion in smaller dimensional latent space. It is to be noted that DiffusionCLIP [40] needs to be fine-tuned for specific text prompts using a set of images ($\sim 30\text{-}50$ images for each prompt), which takes $2\text{--}6$ minutes. Our method also scales well in terms of performing manipulations on multiple images in parallel, in contrast to VQGAN+CLIP, where manipulation on only 2 images could be performed in parallel.

Method	#images	GPU Memory	run-time	(n_{for}, n_{rev})
LDEdit	1	8831MB	$2.02s \pm 5.58$ ms	(25,25)
LDEdit	24	16947MB	22.6 ± 169 ms	(25,25)
LDEdit	1	8831MB	$6.05s \pm 35.6$ ms	(75,75)
LDEdit	24	16947MB	$67.2s \pm 704$ ms	(75,75)
VQGAN+CLIP [19]	1	10413 MB	4-6 mins	—
VQGAN+CLIP [19]	2	18933 MB	5-8 minutes	—
DiffusionCLIP [40]	1	5385MB	$11.54s \pm 66.3$ ms	(200,40)
DiffusionCLIP [40]	1	5385MB	$4.01s \pm 10.5$ ms	(40,40)
DiffusionCLIP [40]	24	15257MB	$156.94s \pm 470$ ms	(200,40)

Table 2: Comparing inference times and GPU memory usage of LDEdit with VQGAN+CLIP [19] and DiffusionCLIP [40]. Images are of dimension 256×256 . n_{for} and n_{rev} refer to the number of forward and reverse diffusion steps. Mean and standard deviation of run-times over 10 runs are reported for LDEdit and DiffusionCLIP.

6 Discussion and Conclusions

We proposed LDEdit, a fast and flexible approach to open domain image manipulation using arbitrary text prompts. Our approach utilizes recent text-to-image latent diffusion model to achieve zero-shot manipulation. Experiments demonstrate that the proposed method can accomplish fast and diverse manipulation making our approach a versatile tool to facilitate efficient user-guided editing. As with other image generation and manipulation methods, there is a potential for LDEdit being misused by bad actors for generating deepfakes and doctored pictures for propaganda. Further, since LDEdit leverages a pretrained text to image latent diffusion model, our approach inherits the inherent biases of its training dataset, including, but not limited to gender, age, and ethnicity of people and cultural biases. a

References

- [1] 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] Rameen Abdal, Peihao Zhu, John Femiani, Niloy J. Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *SIGGRAPH*, 2022.
- [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*, pages 6711–6720, 2021.

- [4] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18511–18521, 2022.
- [5] Amjad Almahairi, Sai Rajeshwar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *International Conference on Machine Learning*, pages 195–204. PMLR, 2018.
- [6] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.
- [7] Omer Bar-Tal, Dolev Ofri-Amar, Rafaail Fridman, Yoni Katen, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Proc. ECCV*, 2022.
- [8] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4502–4511, 2019.
- [9] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021.
- [10] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P Breckon, and Chris G Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. *arXiv preprint arXiv:2111.12701*, 2021.
- [11] Andrew Brock, Theodore Lim, JM Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. In *International Conference on Learning Representations*, 2017.
- [12] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=B1xsqj09Fm>.
- [13] Gerry Chen, Alice Dumay, and Mengyi Tang. diffvg+CLIP: Generating painting trajectories from text. *preprint*, 2021.
- [14] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.
- [15] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.
- [16] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5771–5780, 2020.

- [17] Guillaume Couairon, Asya Grechka, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Flexit: Towards flexible semantic image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18270–18279, June 2022.
- [18] Katherine Crowson. CLIP guided diffusion HQ 256x256. Colab Notebook. URL https://colab.research.google.com/drive/12a_Wrfi2_gwwAuN3VvMTwVMz9TfqctNj.
- [19] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, 2022.
- [20] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- [21] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. CogView: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34, 2021.
- [22] Tan M. Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [23] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5706–5714, 2017.
- [24] Patrick Esser, Robin Rombach, Andreas Blattmann, and Björn Ommer. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in Neural Information Processing Systems*, 34:3518–3532, 2021.
- [25] Patrick Esser, Robin Rombach, and Björn 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.
- [26] Kevin Frans, Lisa B Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint arXiv:2106.14843*, 2021.
- [27] 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. 2022.
- [28] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021.
- [29] Federico A Galatolo, Mario GCA Cimino, and Gigliola Vaglini. Generating images from caption and vice versa via clip-guided generative latent space search. *arXiv preprint arXiv:2102.01645*, 2021.

- [30] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.
- [31] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [32] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3012–3021, 2020.
- [33] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022.
- [34] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.
- [35] Minghui Hu, Yujie Wang, Tat-Jen Cham, Jianfei Yang, and P.N. Suganthan. Global context with discrete diffusion in vector quantised modelling for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11502–11511, June 2022.
- [36] 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.
- [37] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.
- [38] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.
- [39] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.
- [40] Gwanghyun Kim and Jong Chul Ye. Diffusionclip: Text-guided image manipulation using diffusion models. *arXiv preprint arXiv:2110.02711*, 2021.
- [41] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [42] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18062–18071, June 2022.

- [43] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019.
- [44] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020.
- [45] Xihui Liu, Zhe Lin, Jianming Zhang, Handong Zhao, Quan Tran, Xiaogang Wang, and Hongsheng Li. Open-edit: Open-domain image manipulation with open-vocabulary instructions. In *European Conference on Computer Vision*, pages 89–106. Springer, 2020.
- [46] 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.
- [47] Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization. *arXiv preprint arXiv:2112.01573*, 2021.
- [48] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.
- [49] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.
- [50] 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.
- [51] Ryan Murdock. “the big sleep”. URL <https://twitter.com/advadnoun/status/1351038053033406468>.
- [52] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: Manipulating images with natural language. In *Advances in Neural Information Processing Systems*, volume 31, pages 42–51. Curran Associates, Inc., 2018.
- [53] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021.
- [54] 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 39th International Conference on Machine Learning*, volume 162, pages 16784–16804. PMLR, 2022.
- [55] Roni Paiss, Hila Chefer, and Lior Wolf. No token left behind: Explainability-aided image classification and generation. *arXiv preprint arXiv:2204.04908*, 2022.

-
- [56] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Style-CLIP: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021.
 - [57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
 - [58] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.
 - [59] 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.
 - [60] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
 - [61] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1060–1069. PMLR, 2016.
 - [62] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021.
 - [63] 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*, pages 10684–10695, 2022.
 - [64] Chitwan Saharia, William Chan, Huiwen Chang, Chris A Lee, Jonathan Ho, Tim Salimans, David J Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *arXiv preprint arXiv:2111.05826*, 2021.
 - [65] 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.
 - [66] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*, 2021.

- [67] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, number FZJ-2022-00923. Jülich Supercomputing Center, 2021.
- [68] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *Proc. CVPR*, pages 9243–9252, 2020.
- [69] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.
- [70] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.
- [71] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *arXiv preprint arXiv:1907.05600*, 2019.
- [72] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.
- [73] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022.
- [74] Zhicong Tang, Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Improved vector quantized diffusion models. *arXiv preprint arXiv:2205.16007*, 2022.
- [75] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.
- [76] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.
- [77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [78] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.
- [79] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [80] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021.

- [81] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021.
- [82] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. doi: 10.1109/CVPR.2018.00143.
- [83] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jiahui Zhang, Shijian Lu, Miaomiao Cui, Xuansong Xie, Xian-Sheng Hua, and Chunyan Miao. Towards counterfactual image manipulation via clip. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.
- [84] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5908–5916, 2017. doi: 10.1109/ICCV.2017.629.
- [85] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018.
- [86] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [87] Yihao Zhao, Ruihai Wu, and Hao Dong. Unpaired image-to-image translation using adversarial consistency loss. In *European Conference on Computer Vision*, pages 800–815. Springer, 2020.
- [88] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European Conference on Computer Vision*, pages 592–608, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58519-8.
- [89] 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.
- [90] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.
- [91] Junchen Zhu, Lianli Gao, Jingkuan Song, Yuan-Fang Li, Feng Zheng, Xuelong Li, and Heng Tao Shen. Label-guided generative adversarial network for realistic image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [92] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810, 2019.