Prompt Tuning Inversion for Text-Driven Image Editing Using Diffusion Models

Wenkai Dong[1] Song Xue[1] Xiaoyue Duan[1,2*] Shumin Han[1†]
[1]Baidu VIS, [2]Beihang University
{dongwenkai, xuesong06, duanxiaoyue, hanshumin}@baidu.com

Abstract

Recently large-scale language-image models (e.g., text-guided diffusion models) have considerably improved the image generation capabilities to generate photorealistic images in various domains. Based on this success, current image editing methods use texts to achieve intuitive and versatile modification of images. To edit a real image using diffusion models, one must first invert the image to a noisy latent from which an edited image is sampled with a target text prompt. However, most methods lack one of the following: user-friendliness (e.g., additional masks or precise descriptions of the input image are required), generalization to larger domains, or high fidelity to the input image. In this paper, we design an accurate and quick inversion technique, Prompt Tuning Inversion, for text-driven image editing. Specifically, our proposed editing method consists of a reconstruction stage and an editing stage. In the first stage, we encode the information of the input image into a learnable conditional embedding via Prompt Tuning Inversion. In the second stage, we apply classifier-free guidance to sample the edited image, where the conditional embedding is calculated by linearly interpolating between the target embedding and the optimized one obtained in the first stage. This technique ensures a superior trade-off between editability and high fidelity to the input image of our method. For example, we can change the color of a specific object while preserving its original shape and background under the guidance of only a target text prompt. Extensive experiments on ImageNet demonstrate the superior editing performance of our method compared to the state-of-the-art baselines.

1. Introduction

Text-based image editing, a long-standing problem in image processing, aims to modify an input image to align its visual content with the target text prompts. It has drawn increasing attention in recent years and many methods built upon text-to-image generation have been developed. In past years, GAN-based image editing methods [30, 31, 52, 53] achieve impressive results due to the powerful generation abilities of GANs [37, 23, 33, 60]. However, these methods only work well in domains where the models are trained. More recently, diffusion models such as DDPM [18] and score-based generative models [48] have demonstrated competitive or even better capability of generating images compared to VAE-, GAN-, flow- and autoregressive-based models [36, 14, 38, 12]. Especially, large-scale language-image models (LLIMs), such as Imagen [43], DALL-E2 [35] and Stable Diffusion [41], have attracted unprecedented attention from the research community and public society. With the help of large-scale pre-trained language models [34, 10], LLIMs can generate high-fidelity images well aligned with the provided text prompts without further fine-tuning. To fully leverage the generation and generalization capabilities of LLIMs, we aim to develop a text-driven image editing method based on open-sourced LLIMs, *e.g.*, Stable Diffusion [41].

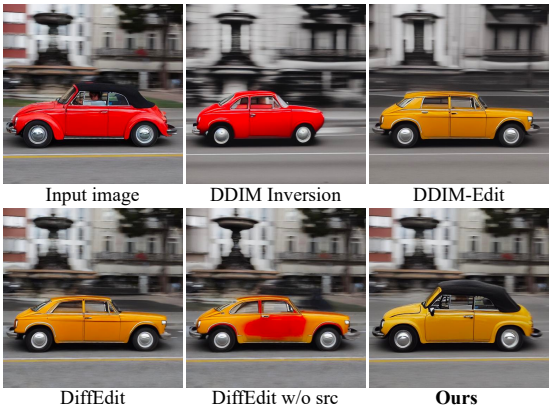

Figure 1. Illustration of different methods in editing the color of the car. Methods based on the original DDIM inversion (*i.e.*, DDIM-Edit, DiffEdit and DiffEdit w/o src) cannot preserve the shape of the car. In contrast, our method successfully changes the color while preserving the structural information. The target text is "a yellow car". The source text is "a red car" for DiffEdit.

Editability and **fidelity** are two essential requirements

*This work was done when Xiaoyue Duan was an intern at Baidu VIS.
†Corresponding author

of image editing tasks. The former requires that the edited images are supposed to contain visual contents well aligned with the corresponding textual contents provided in the target prompts, while the latter expects that areas other than the edited parts should stay as close to those of the input image as possible. For example, when modifying the color of a specific object, its other attributes (*e.g.*, size and shape) are expected to be preserved. As shown in Fig. 1, given an image of a red car and the target text prompt ("a yellow car"), the desired edited image should contain a yellow car while keeping the background as well as the car's size and shape unchanged. To achieve this editing, the simplest way is to first invert the image to a noisy latent via the reversed deterministic DDIM sampling process [46], and then obtain the edited image via the deterministic DDIM sampling process with the guidance of the target prompt embedding. We refer to this approach as "DDIM-Edit" in our paper. Although this approach successfully turns the color of the car to yellow (see "DDIM-Edit" in Fig. 1), the background and the shape of the car change drastically, which obviously fails to meet the requirement of high fidelity. The reason lies in that the deterministic DDIM sampling process cannot be reversed perfectly in practice. A slight error is amplified by a large classifier-free guidance scale, and is accumulated in each sampling step, which consequently results in a significantly different image.

To improve fidelity, some methods consider the image editing tasks as inpainting tasks, which require users to explicitly provide masks of the inpainting regions [3, 25]. With the mask prior, the background can remain the same, but masking out image contents also removes important structural information that is helpful in the editing process, leading to unsatisfactory editing results. Moreover, asking users to provide masks is cumbersome and not suitable for quick and intuitive text-driven image editing. As a solution, DiffEdit [7] presents an algorithm that can automatically generate a mask given a target text prompt to locate the region to be edited. However, the editability of DiffEdit largely depends on DDIM-Edit, which may fail to preserve the structural information of the edited object, *e.g.*, the shape of the car (see "DiffEdit" in Fig. 1). Moreover, to generate an accurate mask, DiffEdit requires a precise text description of the input image (referred to as "source text"), hampering the editing efficiency. Without the source text (see "DiffEdit w/o src" in Fig. 1), the automatically generated mask cannot locate the body of the car accurately, further decreasing editability.

In this work, we aim to propose an image editing method to mitigate all the above problems, *i.e.*, the method should be user-friendly, generalizable to various domains, and generate edited images with high fidelity. Specifically, for a quick and intuitive text-based method, users only need to provide an input image and the corresponding target text prompts, without the need for a mask or a source text describing the input image. Secondly, the method should be able to operate on real images from various domains. Thirdly, the objects should be precisely edited with the background preserved. In some cases, only certain attributes of the objects should be modified, while other attributes are supposed to be left untouched.

To achieve these merits, we believe that image editing needs a new inversion method based on diffusion models to reconstruct the input image. Inspired by the classifier-free guidance [20] and textual-inversion methods [27], we propose a **Prompt Tuning Inversion** method to encode the information of the input image into a conditional embedding. More specifically, we first apply DDIM inversion to the input image latent to obtain a sequence of noisy ones. These noisy latents can be taken as a prior trajectory for reconstructing the original image. Then, we introduce a learnable embedding in the sampling process. The diffusion model reconstructs the input image step by step along the trajectory conditioned on this embedding while optimizing it at the same time. In this way, the contents of the input image are learned in the embedding. Finally, we obtain a new conditional embedding by linearly interpolating between the optimized embedding and the target embedding, resulting in a representation that combines both the structural information of the input image and the visual content of the target text.

Overall, our proposed method consists of two stages. In the first stage, we encode the information of the input image into a learnable conditional embedding via prompt tuning in the reconstruction process. In the second stage, a new conditional embedding is computed by linearly interpolating between the target embedding and the optimized one obtained in the first stage, which boosts a trade-off between editability and fidelity. The classifier-free guidance is then applied to sample the edited image. In sum, our contributions are as follows:

- We propose a user-friendly text-driven image editing method which requires only an input image and a target text for editing, without any need for user-provided masks or source descriptions of the input images.
- We propose a Prompt Tuning Inversion method for diffusion models which can quickly and accurately reconstruct the original image, providing a strong basis for sampling edited images with high fidelity to the inputs.
- We compare against the state-of-the-art methods both qualitatively and quantitatively, and show that our method outperforms these works in terms of the trade-off between editability and fidelity.

2. Related work

Text-to-image synthesis. Text-guided synthesis has been widely adopted for image generation [12, 54, 4, 41]. Works based on generative adversarial networks (GANs) [37, 23, 33, 60] have been proposed for text-to-image synthesis. CLIP-based methods [8, 54] have also been proposed to utilize the language-image priors from a pre-trained CLIP [34] model to generate images from texts. Recently, works [11, 18, 19, 29] based on the Diffusion Probabilistic Models (DPM) [45] have achieved state-of-the-art results in text-to-image synthesis. Among these works, Latent Diffusion Model (LDM) [41] trains DPM in the latent space using a powerful pre-trained auto-encoder, and introduces a cross-attention layer into the model architecture, thus turning the diffusion model into a powerful and flexible generator with greatly improved visual fidelity. Our work of image editing is based on LDM [41] thanks to its powerful image generation capability.

Image editing. Image editing with generative adversarial networks (GANs) has been studied extensively [30, 31, 52, 53]. Some other techniques also leverage the image-text alignment capability of CLIP [34] and transfer it to the framework of GANs [2, 49, 55]. More recently, the development of diffusion models [18, 45, 47] provides a more flexible design space than GANs for the editing task, while following a simpler training setup (*e.g.*, SDEdit [26] and ILVR [6]). Textual Inversion [13] and Dream-Booth [42] demonstrate the capability to generate diverse images with unique object characteristics by fine-tuning the diffusion model with multiple images. Imagic [21] and UniTune [51], which are based on the powerful Imagen model [43], also show impressive editing performance. However, the above methods require restrictive fine-tuning of the pre-trained model, and thus may not fully leverage the generalization ability of the pre-trained model due to overfitting or language drift. Other methods [3, 28] require a user-provided mask to guide the diffusion process, making it hard for them to be interactive. To achieve text-only interactive editing, some optimization-free methods have been proposed recently (*e.g.*, Prompt-to-Prompt [16] and DiffEdit [7]) to automatically infers a mask before editing.

Inversion. In the GAN literature, the inversion process requires one to find a corresponding latent representation of the given image [59, 56]. This process has been extensively studied for GANs [1, 61, 15, 39, 32, 50]. For diffusion models, the inversion requires to find a noise map and a conditional vector corresponding to a generated image but simply adding noise and denoising it may arouse the problem that the image content can be changed drastically. Works [6, 11, 35] have been proposed to improve the inversion process. However, it is still challenging for these methods to generate new instances of a given example while maintaining fidelity. Textual Inversion [13] and Dream-Booth [42] propose to learn concepts from images through textual inversion by either directly optimizing the embedding of the textual concept or fine-tuning the diffusion models, which can be computationally inefficient. Null-Text Inversion [27] modifies the unconditional textual embedding that is used for classifier-free guidance instead of the input text embedding, which enables applying prompt-based editing without the cumbersome tuning of the model parameters. Different from these methods, our method encodes the information of the input image into a learnable conditional embedding, which provides a helpful prior in sampling edited image with high fidelity to the input image.

3. Methodology

Given a real or synthesized image $\mathcal{I}$, we aim to edit $\mathcal{I}$ to get an edited image $\mathcal{I}^*$ with the guidance of text. Different from existing methods which require source prompts provided by users or produced by an off-the-shelf image captioning model, our proposed editing process is guided by only target or edited prompt $\mathcal{P}^*$. An overview of our method is provided in Fig. 2, which consists of two stages. In the first stage, we encode the information of the input image into a learnable conditional embedding via prompt tuning in the reconstruction process. A new conditional embedding is then computed in the second stage by linearly interpolating between the target embedding and the optimized one from the first stage, thus achieving effective editing while maintaining high fidelity. Conditioned on this interpolated embedding, the classifier-free guidance is adopted to sample the final edited image.

3.1. Background and preliminaries

Diffusion models. Diffusion probabilistic models are designed to learn a data distribution by gradually denoising normally distributed noise, which corresponds to learning to reverse a fixed forward diffusion process:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}). \quad (1)$$

In the forward process, normally distributed noise is gradually added to the sample x_{t-1} to obtain a more noisy variant x_t. The noise is dependent on a variance schedule α_t where $t \in 1, ..., T$, with T being the total number of steps, x_0 the original image, and x_T approximately the standard Gaussian noise. The reverse process is defined with parameters θ:

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta). \quad (2)$$

Using a fixed variance Σ_θ, only the mean value $\mu_\theta(x_t, t)$ needs to be learned. With the parameterization trick, the network ϵ_θ is trained to predict the noise ϵ, resulting in a loss, where c represents the conditional embedding:

$$\mathbb{E}_{t,x_0,\epsilon}[||\epsilon - \epsilon_\theta(x_t, t, c)||^2]. \quad (3)$$

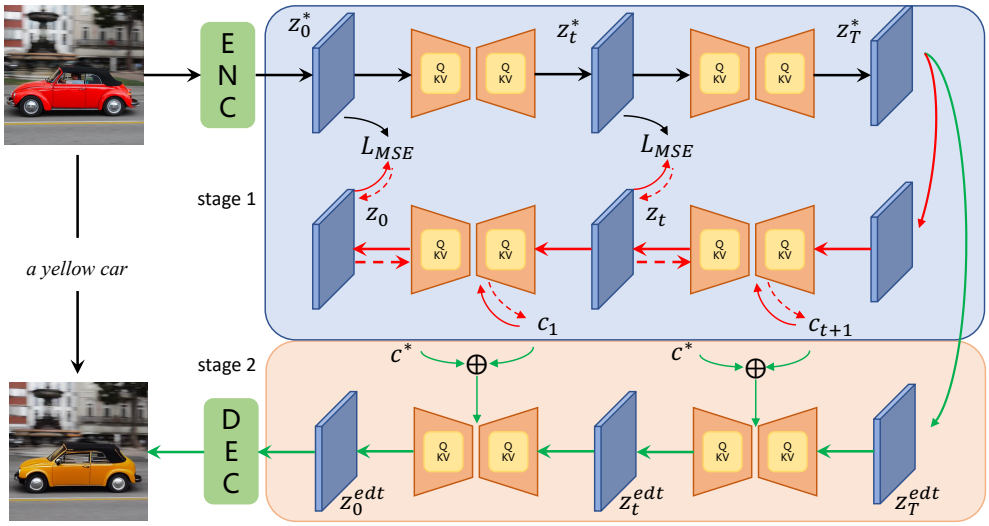

Figure 2. An overview of our proposed image editing method. **Stage 1**: we first apply DDIM inversion to the input image embedding to obtain a diffusion trajectory $\{z_t^*\}_{t=0}^T$. Then we reconstruct the input image along with the reversed trajectory by optimizing the learnable conditional embedding c_t. **Stage 2**: we perform classifier-free guidance sampling conditioned on a linear interpolation between target embedding c^* and c_t at each diffusion step. $\bigoplus$ denotes element-wise weighted addition. Red dashes indicate the gradient flow in PTI.

In this work, we employ the deterministic DDIM sampling [46]:

$$x_{t-1} = \sqrt{\alpha_{t-1}} f_\theta(x_t, t) + \sqrt{1-\alpha_{t-1}} \epsilon_\theta(x_t, t), \quad (4)$$

where f_θ is the prediction of x_0 given x_t at step t. Given a noisy image x_T, the noise is gradually removed to generate an image x_0 by applying Eq. 4 for T steps.

Latent diffusion. Instead of operating in the image pixel space, Latent Diffusion Models [41] (LDMs) utilize an autoencoder to learn a latent space which is perceptually equivalent to the pixel space. First, an encoder ENC is adopted to map a given image x_0 into a latent embedding z_0. Then a decoder DEC is designed to reconstruct the input image given z_0, *i.e.*, $DEC(ENC(x_0)) \approx x_0$. The encoder downsamples the original images by a factor of 4 or 8. In this way, the diffusion model operates on a much smaller representation with lower time complexity and memory burden. Thus, for our method, we apply one of the state-of-the-art LDMs, Stable Diffusion [41]. In the forward and reverse process described above, we only need to replace the image x_t with its latent embedding z_t at each step.

Classifier-free guidance. Our editing method is built upon text-guided diffusion models. In Stable Diffusion, the text $\mathcal{P}$ is fed into a pre-trained CLIP [34] text encoder τ_θ to obtain its corresponding embedding and the underlying UNet model is augmented with the cross attention mechanism, which is effective for generating visual contents conditioned on the text $\mathcal{P}$. One of the key challenges in this kind of generation models is the amplification of the effect induced by the conditional text. To this end, the classifier-free guidance technique is proposed, where the prediction for each step is a combination of conditional and unconditional predictions. Formally, let $c = \tau_\theta(\mathcal{P})$ be the conditional embedding vector and $\varnothing = \tau_\theta(\text{“”})$ be the unconditional one, the classifier-free guidance prediction is calculated by:

$$\tilde{\epsilon}_t = \epsilon_\theta(z_t, t, \varnothing) + \omega \cdot (\epsilon_\theta(z_t, t, c) - \epsilon_\theta(z_t, t, \varnothing)), \quad (5)$$

where ω is the guidance scale parameter.

3.2. Problems of DDIM inversion

Given an input image $\mathcal{I}$ and a target prompt $\mathcal{P}^*$, we aim to edit $\mathcal{I}$ to make its visual content consistent with textual content in $\mathcal{P}^*$, while preserving a maximal amount of details from $\mathcal{I}$. The above two aspects are referred to as **editability** and **input image fidelity**, respectively.

To achieve effective editing while maintaining high fidelity, we first need to inverse the input image into an appropriate noise map, based on which the edited image can be sampled. Eqs. 1 and 2 show a naive way to add noise to the input image and then denoise it through the diffusion network, respectively. However, as the sampling process is stochastic, the samples generated from the same latent can be different every time. Even if the sampling process becomes deterministic, the random noise in the forward process still makes the generated image content change significantly. To address this issue, DiffusionCLIP [22] reverses the deterministic DDIM sampling process in Eq. 4 based on the assumption that the ordinary differential equation (ODE) process can be reversed within the limit of small steps:

$$z_{t+1} = \sqrt{\alpha_{t+1}} f_\theta(z_t, t) + \sqrt{1-\alpha_{t+1}} \epsilon_\theta(z_t, t), \quad (6)$$

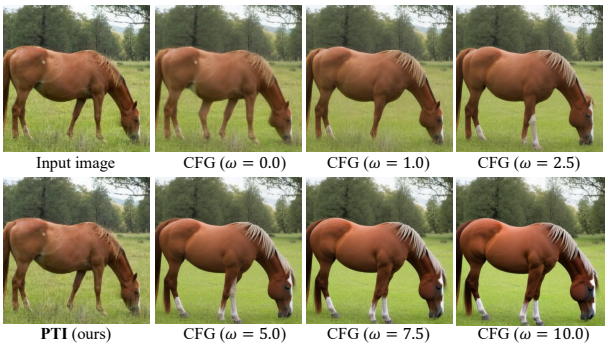

Figure 3. Reconstruction quality of our Prompt Tuning Inversion, and DDIM inversion with different classifier-free guidance (CFG) scales ω in the sampling process. $\omega = 0$ in the forward process for all methods.

ω_{enc} \ ω_{dec}	0.0	1.0	2.5	5.0	7.5
0.0	**21.36**	19.79	17.04	14.88	13.64
1.0	17.16	**21.94**	18.22	15.47	14.02
2.5	14.51	15.73	**18.13**	15.74	14.24
5.0	11.30	11.48	11.70	**12.12**	12.06

Table 1. Reconstruction quality by measuring the PSNR score of DDIM inversion with different classifier-guidance scales ω. ω_{enc} and ω_{dec} denote the guidance scale used in the DDIM forward and sampling processes, respectively.

where z_t is the latent embedding of x_t.

To investigate the reconstruction performance of DDIM, we first invert the latent embedding of the input image into noise maps via Eq. 6 and then use the deterministic DDIM sampling process in Eq. 4 to reconstruct the input. Note that both processes are performed with unconditional diffusion models, *i.e.*, the classifier-free guidance scale in Eq. 5 is set to $\omega = 0$ for both forward and reverse sampling processes. Although a slight error is incorporated in every step as ODE process cannot be reversed perfectly in practice, the accumulated error is negligible, and DDIM inversion can nearly reconstruct the original image (see "CFG ($\omega = 0.0$)" in Fig. 3). However, to generate an image well aligned with the conditional text prompt using Stable Diffusion, a large guidance scale $\omega > 1$ is necessary for the sampling process in Eq. 4. This arouses the problem that when enlarging ω, the generated images are far from the original ones as shown in Fig. 3. We believe that when ω of the sampling process is different from that of the forward process, the accumulated error would be amplified, leading to unsatisfactory reconstruction quality. This can be illustrated in Table 1, *i.e.*, the best reconstruction quality in each line is obtained when ω used in the DDIM sampling process is the same as that used in the forward process. Even if ω used in the forward and sampling processes are kept the same, PSNR still decreases with ω increasing (see the numbers in bold in Table 1). The above analysis demonstrates that it is hard for DDIM inversion to achieve a satisfactory trade-off between editability (which requires larger ω) and fidelity (which requires smaller ω). To address this issue, we propose a new inversion technique, *i.e.*, Prompt Tuning Inversion.

3.3. Prompt tuning for inversion

To successfully invert real images into the model's domain, recent works optimize the textual encoding, the network's parameters, or both. Motivated by Pivotal Inversion [40], we replace the conditional embedding of the text prompt with an optimized one, referred to as **Prompt Tuning** in this work. Namely, for each input image, we optimize only the conditional embedding c so that it encodes important information of the input image which helps the reconstruction. The parameters of the diffusion network and the text encoder τ_θ are frozen during prompt tuning.

We first initialize $z_0^* = z_0 = ENC(x_0)$, and adopt DDIM inversion with $\omega = 0$ to obtain a trajectory of noisy latent codes $\{z_t^*\}_{t=1}^T$. Then we initialize $\tilde{z}_T = z_T^*$ and perform the following optimization to the conditional embedding c with $\omega > 1$ for the timestamps $t = T, ..., 1$, each for N iterations:

$$c_t = \arg\min_{c_t} ||z_{t-1}^* - z_{t-1}(\tilde{z}_t, t, c_t)||^2. \tag{7}$$

For brevity, $z_{t-1}(\tilde{z}_t, t, c_t)$ denotes applying a DDIM sampling step using $\tilde{z}_t$ and the conditional embedding c_t at the timestep t. At the end of each timestep, we update

$$\tilde{z}_{t-1} = z_{t-1}(\tilde{z}_t, t, c_t). \tag{8}$$

Finally, we can reconstruct the input image by using the noise latent $\tilde{z}_T = z_T^*$ and the optimized conditional embeddings $\{c_t\}_{t=1}^T$. In the next subsection, we will introduce the approach to editing images with the target text prompt and the conditional embeddings.

3.4. Prompt tuning for editing

Since the sequence of the conditional embeddings $\{c_t\}_{t=1}^T$ is optimized to fully reconstruct the input image, we believe that these optimized conditional embeddings have contained the most information of the original image, and thus ensure high fidelity. To achieve the desired modification, these optimized embeddings are adopted to perform the editing by advancing in the direction of the target text embedding $c^* = \tau_\theta(\mathcal{P}^*)$ to ensure good editability also. More formally, in the second stage, we simply interpolate between the target embedding c^* and the optimized c_t linearly at each timestamp. For a given hyper-parameter $\eta \in (0, 1]$, we obtain

$$c_t = \eta \cdot c^* + (1 - \eta) \cdot c_t, \tag{9}$$

where the first term ensures the effective editability corresponding to the semantic contents in the target text, while

the second term guarantees a good reconstruction of the original image. The algorithm is presented in Lines 14-21 in Algorithm 1. Note that when $\eta = 0$ or $\eta = 1$, the output of our editing method is the reconstructed original image, or the output of the baseline DDIM-Edit, respectively.

Intuitively, our editing method is to find an intermediate representation between the original image and the output of DDIM-Edit. For a desired modification, the intermediate representation is supposed to contain both the structural information of the source image and the semantic contents of the target text prompt. Eq. 9 is only one way to achieve this, which we refer to as **condition interpolation**. We also test a different interpolation method (referred to as **latent interpolation**), where we linearly interpolate between the noisy latent z_t and z_t^* at each timestamp:

$$z_t^{edt} = \eta \cdot z_t + (1 - \eta) \cdot z_t^*, \tag{10}$$

where z_t^* is the noisy latent calculated by DDIM inversion via Eq. 6, and z_t is the latent obtained in the vanilla DDIM sampling process conditioned on the target embedding. Although this approach is more simple since the process of prompt tuning is no longer needed, we observe that this interpolation method may lead to cluttered images. This is because the interpolation of latent embeddings mixes the source object and the edited object spatially, rather than semantically (as condition interpolation does), leading to cluttered contents in images.

3.5. Discussion

Our proposed image editing method shares similar motivations with existing works [21, 51, 27, 16], all of which aim to modify an image in a text-driven and mask-free manner. However, our approach differs from them significantly in the following aspects:

1) Imagic [21] and UniTune [51] finetune the diffusion models for hundreds of steps to maintain high fidelity to the input image. In contrast, we only need to optimize the conditional embedding, which greatly reduces computational budgets.

2) Our proposed Prompt Tuning Inversion is inspired by the Null-Text Inversion method [27]. However, Null-Text Inversion chooses to optimize the unconditional embedding while we optimize the conditional one. Moreover, the editing process of Null-Text Inversion is achieved by the cross-attention map control in Prompt-to-Prompt [16], which requires an additional description of the input image. Compared to theirs, our method only needs the target text prompt and is thus more user-friendly.

Algorithm 1: Prompt Tuning Inversion for Editing

Input: An input image $\mathcal{I}$ and a target prompt embedding $c^* = \tau_\theta(\mathcal{P}^*)$
Output: Edited image $\mathcal{I}^*$.

1 // DDIM Inversion
2 Set guidance scale $\omega = 0$, $z_0^* = ENC(\mathcal{I})$;
3 Compute the intermediate trajectory $\{z_t^*\}_{t=0}^T$ using DDIM inversion over $\mathcal{I}$ without conditional guidance via Eq. 6;
4 // Prompt Tuning
5 Set guidance scale $\omega > 1$, $\eta \in (0, 1]$;
6 Initialize $\tilde{z}_T \leftarrow z_T^*$, $c_T \leftarrow c^*$;
7 **for** *t = T, T-1, ..., 1* **do**
8 **for** *j = 0, ..., N-1* **do**
9 $c_t \leftarrow c_t - \beta \nabla_c ||z_{t-1}^* - z_{t-1}(\tilde{z}_t, t, c_t)||_2^2$;
10 **end**
11 $\tilde{z}_{t-1} \leftarrow z_{t-1}(\tilde{z}_t, t, c_t)$, $c_{t-1} \leftarrow c_t$
12 **end**
13 // Editing
14 Set $z_T^{edt} \leftarrow z_T^*$;
15 **for** *t = T, T-1, ..., 1* **do**
16 $c_t = (1 - \eta) \cdot c_t + \eta \cdot c^*$;
17 $z_{t-1}^{edt} = z_{t-1}(z_t^{edt}, t, c_t)$;
18 **end**
19 $\mathcal{I}^* = DEC(z_0^{edt})$;
20 return $\mathcal{I}^*$

4. Experiments

4.1. Setup

Implementation details. In our experiments, we adopt the text-conditional Latent Diffusion Model [41] (known as Stable Diffusion) with 890M parameters trained on LAION-5B [44] at 512×512 resolution. For the DDIM schedule, we adopt 50 steps and retain the original hyper-parameter choices of Stable Diffusion. The encoding ratio parameter is set to 0.8. The number of iterations to optimize c per diffusion step is set to $N = 1$. The hyper-parameters β and η in Algorithm 1 are set to 0.1 and 0.9, respectively, unless specified. These allow editing an image in ~ 1 minute on a single Tesla V100 GPU. For better performance, we adopt attention maps to localize the edited regions (referred to as local blending), and re-weight the attention maps as in [16].

Evaluation and datasets. In semantic image editing, the visual content of the edited image is supposed to align well with the target text prompt (editability) while staying close to the input image in terms of the unedited parts (fidelity). For a given method, better editability usually comes at the cost of decreased fidelity to the input image, and vice versa. This forms a trade-off curve between the two objectives.

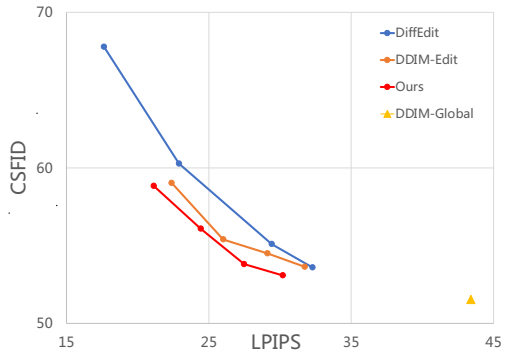

Figure 4. Comparison with DiffEdit and DDIM-Edit on ImageNet. For all methods, we set the DDIM encoding ratio to 0.8, and only vary the mask threshold to draw the trade-off curve.

Following DiffEdit [7], we evaluate different editing methods by comparing their trade-off curves on ImageNet [9]. Specifically, given an image of one class from ImageNet, we aim to edit it to another class of ImageNet as instructed by the target text prompt. The editability and fidelity are measured using the LPIPS perceptual distance [58] and CSFID, which is a class-conditional FID metric [17], respectively. The former measures the distance with the input image while the latter measures both image realism and consistency *w.r.t.* the target class. For both metrics, lower values indicate better editing performance.

4.2. Comparison with other methods on ImageNet

We compare our method with DiffEdit and our baseline DDIM-Edit, since they both share the same DDIM forward process and a similar sampling process. Besides, they load the same publicly available pre-trained weights for a fair comparison. To leverage the generalization capability of large-scale language-image models, we adopt the text-conditional Stable Diffusion model "sd-v1-4" instead of the class-conditional model trained on ImageNet as the pre-trained model.

As pointed out by DiffEdit [7], different editing methods often have hyper-parameters which control editability, *e.g.*, the mask threshold, or the encoding ratio. A lower mask threshold or higher encoding ratio leads to stronger editing. In our proposed method, we can also control the editing strength by varying the conditional interpolation ratio η introduced in Eq. 9. In our evaluation, we fix the DDIM encoding ratio and draw the trade-off curve by varying the mask threshold for all methods[1]. The results are presented in Fig. 4, where "DDIM-Global" denotes that the images are edited via "DDIM-Edit" but without using any masks, *i.e.*, the editing is performed globally. This can be regarded as a lower bound of fidelity for all methods. As the mask threshold increases, LPIPS decreases since fewer parts of the image are edited. Compared to DiffEdit, our baseline method DDIM-Edit achieves a better trade-off. Note that the only difference between the two methods is the approach to generating masks. The comparison shows that inferring editing masks using cross-attention maps, as adopted by DDIM-Edit, is more appropriate. Based on DDIM-Edit, our method can further improve the fidelity to the input images while maintaining editability. The best CSFID-LPIPS trade-off of our method demonstrates its superiority over DiffEdit and the baseline.

[1]As there was no official implementation of DiffEdit available at the time of writing, we adopted the unofficial implementation for inferring editing masks from https://github.com/LuChengTHU/dpm-solver.

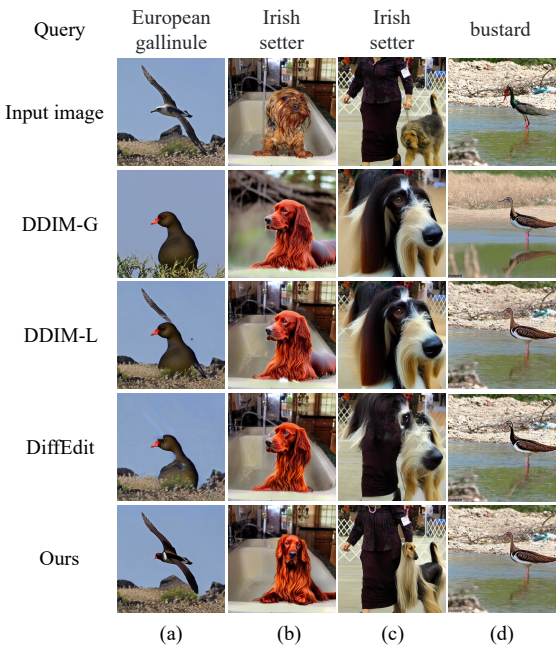

Figure 5. Editing examples on ImageNet by our method and other methods. DDIM-G/DDIM-L indicates the baseline method DDIM-Edit with/without the local blend trick.

We also present qualitative examples of these methods. As shown in Fig. 5, without automatically generated masks, "DDIM-G" tends to modify images globally. For simple cases (*e.g.*, example (d)), image editing methods with the original DDIM inversion works well. However, for complex cases, we observe undesired and unreasonable edits to the objects. In contrast, with the help of the learnable conditional embedding, our method achieves realistic editing while successfully preserving the original details.

4.3. Ablation study

Comparison to existing inversion methods. We randomly select 128 images and their corresponding captions from the COCO validation set [24]. We then apply the following reconstruction methods to each image-caption pair: **(1) AE** denotes the variational auto-encoder with a slight KL-

Method	AE	DDIM	NTI	**PTI (ours)**
PSNR	26.22	13.64	24.45	25.71
SSIM	0.8564	0.4641	0.8270	0.8501

Table 2. Reconstruction quality of different methods measured by PSNR and SSIM. For both metrics, higher values indicate better quality.

iters	1	2	3	4	5
learning rate $\beta = 0.01$					
NTI	17.05	20.12	22.25	23.61	24.45
PTI	19.36	23.20	24.86	25.47	25.71
learning rate $\beta = 0.1$					
NTI	23.08	24.74	25.62	25.82	25.91
PTI	24.74	25.23	25.78	25.90	25.97

Table 3. PSNR scores under different optimizing settings.

penalty used in Stable Diffusion. An image is first encoded by the encoder of AE. Afterwards, the decoder directly reconstructs the image from the latent. Therefore, we consider AE as an upper bound of reconstruction quality. **(2) DDIM** denotes the DDIM inversion method, which is a baseline inversion method. As analyzed in Sec. 3.2, it usually outputs a low-quality reconstruction result under a large classifier-free guidance scale ω. **(3) NTI** denotes the Null-Text Inversion method [27], which is our main point of comparison. Different from our method, it optimizes the unconditional embedding. **(4) PTI** denotes our proposed Prompt Tuning Inversion method. We introduce a learnable conditional embedding and the optimizing details are presented in Algorithm 1.

The experimental results are provided in Table 2. For the diffusion-based inversion methods, we apply the diffusion model in an unconditional manner (*i.e.*, the classifier guidance scale $\omega = 0$) for the DDIM forward process. For the sampling process, we set $\omega = 7.5$. As shown in Table 2, DDIM inversion fails to reconstruct the original images since ω in the sampling process is different from that in the forward process, leading to a low PSNR score (13.64). For NTI and PTI, we set the number of iterations N to 5 and the learning rate β to 0.01. We observe that both methods can reconstruct the images but the reconstruction quality of our method is better (25.71 vs. 24.45). To further demonstrate the effectiveness of our method, we vary N from 1 to 5 and increase the learning rate β from 0.01 to 0.1. The experimental results in Table 3 shows that the reconstruction quality of PTI is always better than NTI under all settings, demonstrating that our method converges faster.

Influence of the hyper-parameter η. We also perform ablation on the hyper-parameter η and the results are presented in Fig. 6. By adjusting η, we can achieve a good trade-off between fidelity and editability. More quantitative analyses are presented in supplementary materials.

Figure 6. Qualitative examples for the effect of η.

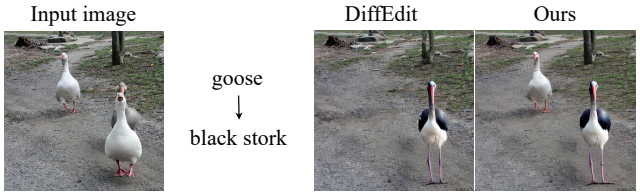

Figure 7. A failure example on ImageNet: when multiple objects exist, only one of them is edited successfully.

5. Conclusion and future work

We propose an intuitive and user-friendly text-based image editing method, which benefits from the superior generation and generalization capacities of large-scale image language diffusion models (*e.g.*, Stable Diffusion). The key idea of our method is that important structural information of the input image can be encoded into conditional embeddings, which can guide the diffusion model to reconstruct the original image via the sampling process. Based on this, our method consists of two stages. In the first reconstruction stage, we propose a novel Prompt Tuning Inversion method which encodes image information to learnable conditional embeddings quickly and accurately. In the second editing stage, we introduce an interpolation which linearly combines the target text embedding with the optimized embedding obtained in the first stage. In this way, the new conditional embedding contains both information from the input image and the target text prompt, resulting in an edited image with an appropriate trade-off between editability and fidelity.

While our method works well in most scenarios, it still faces some limitations. As shown in Fig. 7, there are multiple objects in the input images. However, neither DiffEdit nor our method changes all "gooses" to "black storks". This limitation can possibly be mitigated by operating the attention maps more precisely [5] or adding different modes of conditional control [57], providing a research direction for image editing. Besides, the proposed method still requires running inversion every time given a new text since initialization is based on the target text, and not using attention is not always an advantage. We leave these options for future work.

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. 3

[2] Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9, 2022. 3

[3] 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. 2, 3

[4] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 3

[5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *arXiv preprint arXiv:2301.13826*, 2023. 8

[6] 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. 3

[7] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2, 3, 7

[8] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 88–105. Springer, 2022. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. of CVPR*, 2009. 7

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[12] 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. 1, 3

[13] 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. 3

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[15] 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. 3

[16] 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. 3, 6

[17] 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. 7

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 3

[19] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022. 3

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[21] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 3, 6

[22] 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. 4

[23] Jiadong Liang, Wenjie Pei, and Feng Lu. Cpgan: Content-parsing generative adversarial networks for text-to-image synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 491–508. Springer, 2020. 1, 3

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7

[25] 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. 2

[26] 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*, 2021. 3

[27] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real

images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 2, 3, 6, 8

[28] 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. 3

[29] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3

[30] 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. 1, 3

[31] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 1, 3

[32] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14104–14113, 2020. 3

[33] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514, 2019. 1, 3

[34] 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. 1, 3, 4

[35] 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, 3

[36] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 1

[37] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 1, 3

[38] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 1

[39] 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. 3

[40] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 5

[41] 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. 1, 3, 4, 6

[42] 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. 3

[43] 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, 3

[44] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 6

[45] 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. 3

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 4

[47] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 3

[48] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[49] David Stap, Maurits Bleeker, Sarah Ibrahimi, and Maartje Ter Hoeve. Conditional image generation and manipulation for user-specified content. *arXiv preprint arXiv:2005.04909*, 2020. 3

[50] 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. 3

[51] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2022. 3, 6

[52] Yael Vinker, Eliahu Horwitz, Nir Zabari, and Yedid Hoshen. Image shape manipulation from a single augmented training sample. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13769–13778, 2021. 1, 3

[53] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8798–8807, 2018. 1, 3

[54] Zihao Wang, Wei Liu, Qian He, Xinglong Wu, and Zili Yi. Clip-gen: Language-free training of a text-to-image generator with clip. *arXiv preprint arXiv:2203.00386*, 2022. 3

[55] 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. 3

[56] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[57] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 8

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[59] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 597–613. Springer, 2016. 3

[60] 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. 1, 3

[61] Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. Improved stylegan embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*, 2020. 3