Smooth Diffusion: Crafting Smooth Latent Spaces in Diffusion Models

Jiayi Guo[1,2*], Xingqian Xu[1,3*], Yifan Pu[2], Zanlin Ni[2], Chaofei Wang[2], Manushree Vasu[1], Shiji Song[2], Gao Huang[2†], Humphrey Shi[1,3†]

[1]SHI Labs @ Georgia Tech & UIUC [2]Tsinghua University [3]Picsart AI Research (PAIR)

https://github.com/SHI-Labs/Smooth-Diffusion

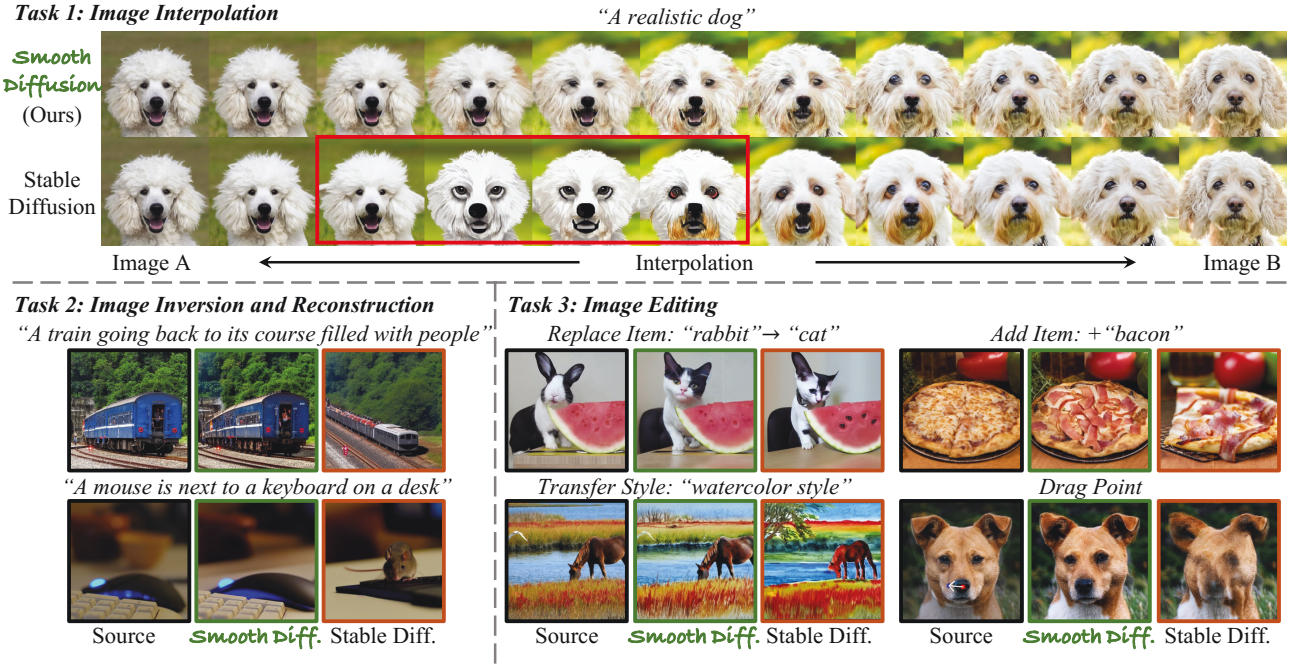


Figure 1. **Smooth Diffusion for downstream image synthesis tasks.** Our method formally introduces **latent space smoothness** to diffusion models like Stable Diffusion [62]. This smoothness dramatically aids various tasks in: 1) improving continuity of transitions in image interpolation, 2) reducing approximation errors in image inversion, & 3) better preserving unedited contents in image editing.

Abstract

Recently, diffusion models have made remarkable progress in text-to-image (T2I) generation, synthesizing images with high fidelity and diverse contents. Despite this advancement, ***latent space smoothness*** *within diffusion models remains largely unexplored. Smooth latent spaces ensure that a perturbation on an input latent corresponds to a steady change in the output image. This property proves beneficial in downstream tasks, including image interpolation, inversion, and editing. In this work, we expose the non-smoothness of diffusion latent spaces by observing noticeable visual fluctuations resulting from minor latent variations. To tackle this issue, we propose* ***Smooth Diffusion****, a new category of diffusion models that can be simultaneously high-performing and smooth. Specifically, we introduce* ***Step-wise Variation Regularization*** *to enforce the proportion between the variations of an arbitrary input latent and that of the output image is a constant at any diffusion training step. In addition, we devise an interpolation standard deviation (ISTD) metric to effectively assess the latent space smoothness of a diffusion model. Extensive quantitative and qualitative experiments demonstrate that Smooth Diffusion stands out as a more desirable solution not only in T2I generation but also across various downstream tasks. Smooth Diffusion is implemented as a plug-and-play Smooth-LoRA to work with various community models. Code is available at https://github.com/SHI-Labs/Smooth-Diffusion.*

*Equal contribution.
†Corresponding authors.

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.00721

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.00721

1. Introduction

In recent years, diffusion models [12, 24, 62] have rapidly grown into very powerful tools for generative AI, particularly for text-to-image generation. The remarkable ability of diffusion models, generating high-quality photorealistic images from open-book contexts, has been highlighted in many research and commercial products. Such success has also inspired various diffusion-based downstream tasks, including image interpolation [29, 78], inversion [13, 44, 54, 71, 77], editing [21, 42, 45, 55, 69, 75, 81, 82], *etc*.

Despite the great success in the generation field, diffusion models occasionally produce low-quality results with undesirable and unpredictable behaviors. Specifically speaking, for image interpolation, the Stable Diffusion Walk (SDW) [29] test examines latent space with spherical linear interpolations, usually resulting in highly fluctuated outputs with unpredictable visual appearance. Examples can be found in Fig. 1 Task 1, in which such interpolation exhibits undesired sharp changes as well as "cartoonization" on photorealistic dog images, highlighted in the red box. For the image inversion task shown in Fig. 1 Task 2, a naive application of DDIM inversion [71] cannot reconstruct images faithfully from the sources. Instead, it generates incorrect colors and object orientations, and misinterprets the computer mouse as an animal mouse. For the image editing task shown in Fig. 1 Task 3, one may notice that only minor text prompt editing can lead to major updates on image contents and layouts, in which the object (*i.e*. the cat's pose, the horse's location, the shape of the pizza) can be wildly and incorrectly altered. Moreover, current diffusion models are unsuited to drag-based editing [69] because a fine-engineered drag method still has a noticeably large chance of breaking objects' shape and semantics.

In this work, we step into an important but underexplored area: to improve the latent space smoothness of diffusion models. Our motivation to enhance latent smoothness comes from the real-world demand to improve the output qualities of the aforementioned downstream tasks. A smooth latent space implies a robust visual variation under a minor latent change. Therefore, enhancing such smoothness could help improve the continuity of image interpolation, expand the capacity of image inversion, and maintain correct semantics in image editing. Notably, prior works in GANs [32, 33, 68] have demonstrated that the smooth latent space of the generator can significantly improve downstream tasks' quality, offering additional evidence of the importance of this area.

To achieve our goal, we propose **Smooth Diffusion**, a new category of diffusion models that can be simultaneously high-performing and smooth. We start our exploration by first formalizing the objective for Smooth Diffusion, in which fixed-size perturbations on a latent noise should produce smooth visual changes on the synthetic image , rounded to a constant ratio . Although one may think that according to the formulation, the smoothness constraint could be an accessible train-time loss. Actually, there is no direct application of such regularization from inference to training, and the challenge lies in the fact that in each training iteration (*i.e*., back-propagation), diffusion models optimize only a " -step snapshot" instead of the entire -step diffusion process.

Therefore, we introduce **Step-wise Variation Regularization**, a novel regularization that seamlessly incorporates our Smooth Diffusion's inference-time objective to training. This regularization aims to bound the 2-norm of output variation given a fixed-size change in input at an arbitrary step . The rationale of the reformulation is intuitive: If and exhibit smooth changes at any , then the relation between the latent noise (*i.e*.) and is just the accumulation of smooth variations and thus can be smooth as well. More details can be found in Sec. 3.

In practice, our Smooth Diffusion is trained on top of a well-known text-to-image model: Stable Diffusion [62]. We examine and demonstrate that Smooth Diffusion dramatically improves the latent space smoothness over its baseline. Meanwhile, we conduct extensive research across numerous downstream tasks, including but not limited to image interpolation, inversion, editing, *etc*. Both qualitative and quantitative results support our conclusion that Smooth Diffusion can be the next-gen high-performing generative model not only for the baseline text-to-image task but across various downstream tasks.

2. Related Work

Diffusion models are initiated from a family of prior works including but not limited to [8, 66, 70, 76]. Since then, DDPM [24] introduced an image-based noise prediction model, becoming one of the most popular image generation research. Later works [12, 48, 71] extended DDPM, demonstrating that diffusion models perform on-par and even surpass GAN-based methods [15, 30–33]. Recently, generating images from text prompts (T2I) become an emerging field [11, 19, 28, 47, 62], among which diffusion models [16, 49, 59, 62, 64] have become quite visible to the public. For example, Stable Diffusion (SD) [62] consists of VAE [36] and CLIP [58], diffuses latent space, and yields an outstanding balance between quality and speed. Following SD [62], researchers also explored diffusion approaches for controls such as ControlNet [14, 26, 46, 57, 80, 85, 86, 89–91, 95] and multimodal such as Versatile Diffusion [6, 41, 73, 88]. Works from a different track reduce diffusion steps to improve speed [5, 34, 39, 43, 65, 72, 92, 96], or restrict data and domain for few-shot learning [20, 25, 40, 63], all had successfully maintained a high output quality.

Smooth latent space was one of the prominent proper-

ties of SOTA GAN works [9, 31–33], while exploring such property went through the decade-long GAN research [3, 15], whose goals were mainly robust training. Ideas such as Wasserstein GAN [4, 17] had proved to be effective, which enforced the Lipschitz continuity on discriminator via gradient penalties. Another technique, namely path length regularization, related to the Jacobian clamping in [51], was adapted in StyleGAN2 [32] and later became a standard setting for GAN-based generators [10, 37, 87, 94]. Benefiting from the smoothness property, researchers managed to manipulate latent space in many downstream research projects. Works such as [7, 50, 68, 83] explored latent space disentanglement. GAN-inverse [1, 2, 52, 84] had also proved to be feasible, along with a family of image editing approaches [18, 53, 56, 60, 61, 74, 97]. As aforementioned, our work aims to investigate the latent space smoothness for diffusion models, which by far remains unexplored.

3. Methodology

In this section, we first introduce preliminaries of our method, including diffusion process [24], diffusion inversion [12, 44, 71] and low-rank adaptation [25] (Sec. 3.1). Then Smooth Diffusion is proposed with its definition, objective (Sec. 3.2) and regularization function (Sec. 3.3).

3.1. Preliminaries

Diffusion process [24] is a kind of Markov chain that gradually adds random noise $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{I})$ to ground truth signal $\boldsymbol{x}_0 \sim p(\boldsymbol{x}_0)$, making $\boldsymbol{x}_T$ in a total of T steps. At each step, The noisy data $\boldsymbol{x}_t$ is computed as:

$$\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}_t, \quad t = 1, 2, \cdots, T, \tag{1}$$

where β_t is the preset diffusion rate at step t. By making $\alpha_t = 1 - \beta_t$, $\overline{\alpha_t} = \prod_{t=1}^{T} \alpha_t$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{I})$, we have the following equivalents:

$$\begin{aligned} \boldsymbol{x}_t &= \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t \\ &= \sqrt{\overline{\alpha_t}}\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha_t}}\boldsymbol{\epsilon}, \quad t = 1, 2, \cdots, T. \end{aligned} \tag{2}$$

A diffusion model $\epsilon_\theta(\boldsymbol{x}_t, t)$ is then trained to estimate $\boldsymbol{\epsilon}_t$ from $\boldsymbol{x}_t$, by which one can predict the original signal $\boldsymbol{x}_0$ by gradually remove noise from the degraded $\boldsymbol{x}_T$ [71]. This is commonly known as the backward diffusion process:

$$\widehat{\boldsymbol{x}_{t-1}} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\widehat{\boldsymbol{x}_t} + \sqrt{\alpha_{t-1}}\left(\sqrt{\frac{1}{\alpha_{t-1}}-1} - \sqrt{\frac{1}{\alpha_t}-1}\right)\cdot \epsilon_\theta(\widehat{\boldsymbol{x}_t}, t). \tag{3}$$

Diffusion inversion [12, 44, 71] targets to recover the exact backward diffusion process (*i.e.* $\widehat{\boldsymbol{x}_t}, \epsilon_\theta(\widehat{\boldsymbol{x}_t}, t), t = 1, ..., T$) from a known final prediction $\widehat{\boldsymbol{x}_0}$. One of the common technique for such inversion is *DDIM inversion* [12, 71], which reverses Eq. (3) under a local linear approximation:

$$\widetilde{\boldsymbol{x}_{t+1}} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}\widetilde{\boldsymbol{x}_t} + \sqrt{\alpha_{t+1}}\left(\sqrt{\frac{1}{\alpha_{t+1}}-1} - \sqrt{\frac{1}{\alpha_t}-1}\right)\cdot \epsilon_\theta(\widetilde{\boldsymbol{x}_t}, t), \tag{4}$$

where $\widetilde{\boldsymbol{x}_t}$ represent the estimated $\widehat{\boldsymbol{x}_t}$ at time t. However, DDIM inversion is only a rough estimation. For text-to-image diffusion, a more advanced technique, *Null-Text Inversion* [44], optimizes additional null-text embeddings $\{\varnothing_t\}_{t=1}^{T}$ for each step t, simulating the backward process with $\epsilon_\theta(\boldsymbol{x}_t, t, \xi, \varnothing_t)$, where ξ is the input text embedding. The predicted null-text $\varnothing_t$ is the null input of the classifier-free guidance [23] with a guidance scale w:

$$\epsilon_\theta(\boldsymbol{x}_t, t, \xi, \varnothing_t) = w \cdot \epsilon_\theta(\boldsymbol{x}_t, t, \xi) + (1-w)\cdot \epsilon_\theta(\boldsymbol{x}_t, t, \varnothing_t). \tag{5}$$

Low-rank adaptation (LoRA) [25] is initially proposed to efficiently adapt large pretrained models to downstream tasks. The key assumption of LoRA is that the weight changes required during adaptation maintain a low rank. Given a pretrained model weight $W_0 \in \mathbb{R}^{d\times k}$, its updated weight ΔW is expressed as a low rank decomposition:

$$W_0 + \Delta W = W_0 + BA, \tag{6}$$

where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$ and $r \ll \min(d, k)$. During adaptation, W_0 is frozen, while B and A are trainable.

3.2. Smooth Diffusion

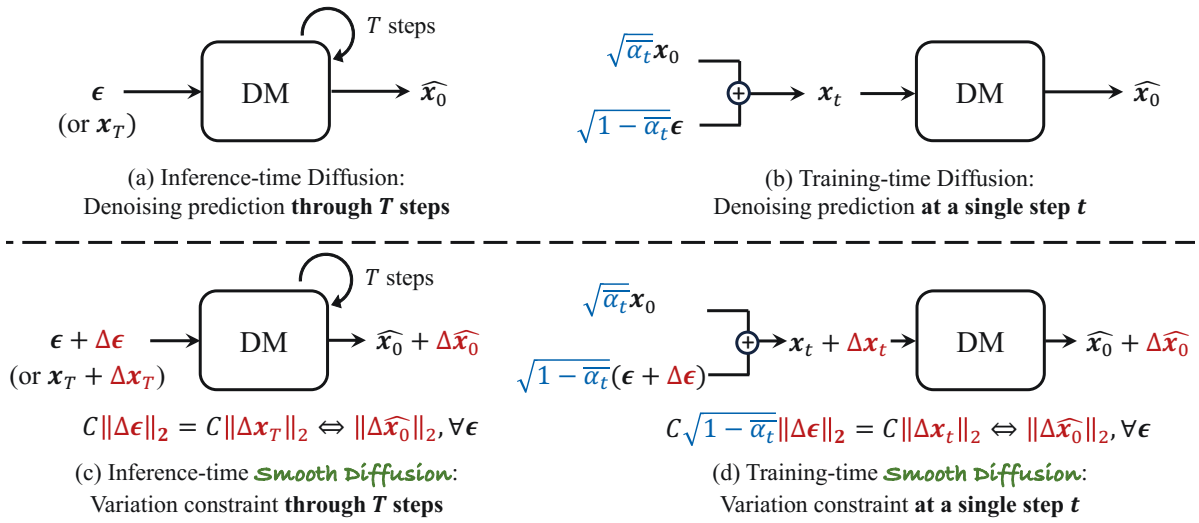
As previously mentioned, modern diffusion models (DM) do not guarantee latent space smoothness, creating not only research gaps between GANs and diffusions but also unexpected challenges in downstream tasks. To address these issues, we propose **Smooth Diffusion**, a novel class of high-performing diffusion models with enhanced smoothness over its latent space. The underlining of Smooth Diffusion is the newly proposed training scheme in which we carried out a **Step-wise Variation Regularization** to enhance model smoothness.

To better explain our aims, we adopt the same terminologies from the standard inference-time diffusion process (Fig. 2a), involving a T steps procedure that transforms the random noise $\boldsymbol{\epsilon}$ (*i.e.*, $\boldsymbol{x}_T$) to the prediction $\widehat{\boldsymbol{x}_0}$. The overall objective of Smooth Diffusion can then be written in Eq. 7: in which we expect that a fixed-size change $\Delta\boldsymbol{\epsilon}$ on $\boldsymbol{\epsilon}$ (*i.e.*, $\Delta\boldsymbol{x}_T$ on $\boldsymbol{x}_T$) will finally lead to a non-zero, fixed-size change $\Delta\widehat{\boldsymbol{x}_0}$ on $\widehat{\boldsymbol{x}_0}$, up to a constant ratio C:

$$\|\Delta\widehat{\boldsymbol{x}_0}\|_2 \Leftrightarrow C\|\Delta\boldsymbol{x}_T\|_2 = C\|\Delta\boldsymbol{\epsilon}\|_2, \ \forall\boldsymbol{\epsilon}, \tag{7}$$

Notice that by definition, $\boldsymbol{x}_T$ is the initial input of the backward diffusion loop in Eq. 3. Since $\boldsymbol{x}_T$ is close to $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{1})$, for simplicity, we make them equivalent in all the following equations.

Nevertheless, one may notice that our inference-time objective in Eq. 7 cannot be directly transformed into a train-



(a) Inference-time Diffusion: Denoising prediction **through T steps**

(b) Training-time Diffusion: Denoising prediction **at a single step t**

(c) Inference-time Smooth Diffusion: Variation constraint **through T steps**

(d) Training-time Smooth Diffusion: Variation constraint **at a single step t**

Figure 2. **Illustration of Smooth Diffusion.** Smooth Diffusion (c) enforces the ratio between the variation of the input latent ($\|\Delta\epsilon\|_2$ or $\|\Delta x_T\|_2$) and the variation of the output prediction ($\|\Delta\hat{x}_0\|_2$) is a constant C. Training-time Diffusion (b) optimizes a "t-step snapshot" of the denoising prediction process in Inference-time Diffusion (a). Similarly, we propose Training-time Smooth Diffusion (d) to optimize a "t-step snapshot" of the variation constraint in Inference-time Smooth Diffusion (c). DM: Diffusion model.

ing loss function. This is because, in one training iteration (*i.e.*, back-propagation), diffusion models optimize only a "t-step snapshot" of the diffusion process (Fig. 2b), where t is uniformly sampled from 1 to T. Hence, the proposed "global" objective (Eq. 7) for the entire T-step process is not accessible in training. Therefore, we need to reformulate our global objective into a **step-wise objective** shown in Eq. 8, which can later be integrated into the diffusion training process as a loss function:

$$\|\Delta\hat{x}_0\|_2 \Leftrightarrow C\|\Delta x_t\|_2 = C\sqrt{1-\bar{\alpha}_t}\|\Delta\epsilon\|_2, \ \forall\epsilon, \quad (8)$$

where C is a non-zero constant. This step-wise objective indicates that at each training step, variations $\Delta\epsilon$ on ϵ should imply variations Δx_t on x_t with a ratio proportional to $\sqrt{1-\bar{\alpha}_t}$. The rationale of Eq. 8 is intuitive: If x_t and $\hat{x}_0$ show smooth changes at any t, then the relation between the latent noise ϵ (*i.e.* x_T) and $\hat{x}_0$ is just the accumulation of smooth variations and thus can be smooth as well.

3.3. Step-wise Variation Regularization

While the motivation and formulation of the Smooth Diffusion objective are presented, how to realize such an objective remains unexplained. Therefore, in this section, we introduce **Step-wise Variation Regularization** to effectively integrate the step-wise objective into diffusion training.

We draw inspiration from the regularization techniques [32, 51] adopted in GAN training. The core idea of Step-wise Variation Regularization is to bound the Jacobian matrix $\mathbf{J}_\epsilon = \partial\hat{x}_0/\partial\epsilon$ of the diffusion system by minimizing the following regularization loss at any x_0, ϵ, and step t:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\Delta\hat{x}_0,\epsilon}\left(\sqrt{1-\bar{\alpha}_t}\|\mathbf{J}_\epsilon^{\mathrm{T}}\Delta\hat{x}_0\|_2 - a\right)^2, \quad (9)$$

where $\Delta\hat{x}_0$ is the normally sampled pixel intensities normalized to unit length, ϵ is a normally sampled noise in Eq. 2, and a is the exponential moving average of $\sqrt{1-\bar{\alpha}_t}\|\mathbf{J}_\epsilon^{\mathrm{T}}\Delta\hat{x}_0\|_2$ computed online during training. In practice, we compute Eq. 9 via standard backpropagation with the following identity:

$$\sqrt{1-\bar{\alpha}_t}\|\mathbf{J}_\epsilon^{\mathrm{T}}\Delta\hat{x}_0\|_2 = \|\nabla_\epsilon(\sqrt{1-\bar{\alpha}_t}\hat{x}_0 \cdot \Delta\hat{x}_0)\|_2. \quad (10)$$

The identity holds since $\Delta\hat{x}_0$ is independently sampled, and uncorrelated with ϵ.

Next, we prove that the proposed objective in Eq. 9 exactly matches our optimization goal in Eq. 8. One preliminary result, proven in [32], is that in high dimensions, Eq. 9 is minimized when $\mathbf{J}_\epsilon$ is orthogonal at any ϵ up to a global scaling factor $\mathcal{K}$ (*i.e.* $\mathbf{J}_\epsilon \cdot \mathbf{J}_\epsilon^{\mathrm{T}} = \mathcal{K} \cdot \boldsymbol{I}$). By applying the orthogonality of $\mathbf{J}_\epsilon$, we have the following:

$$\mathbf{J}_\epsilon^{\mathrm{T}}\Delta\hat{x}_0 = \mathcal{K}\mathbf{J}_\epsilon^{-1}\Delta\hat{x}_0 = \mathcal{K}\frac{\partial\epsilon}{\partial\hat{x}_0} \cdot \Delta\hat{x}_0 = \mathcal{K}\Delta\epsilon. \quad (11)$$

When $\mathcal{L}_{\text{reg}}$ in Eq. 9 reaches its optimal, we then have:

$$a = \sqrt{1-\bar{\alpha}_t}\|\mathbf{J}_\epsilon^{\mathrm{T}}\Delta\hat{x}_0\|_2 = \sqrt{1-\bar{\alpha}_t}\mathcal{K}\|\Delta\epsilon\|_2. \quad (12)$$

Notice that $a = a\|\Delta\hat{x}_0\|_2$, since $\|\Delta\hat{x}_0\|_2 = 1$ is the aforementioned random unit length vector. Hence, we can finally reformulate the expression:

$$\begin{aligned}\|\Delta\hat{x}_0\|_2 &= \frac{\mathcal{K}}{a}\sqrt{1-\bar{\alpha}_t}\|\Delta\epsilon\|_2 \\ &= \mathcal{C}\sqrt{1-\bar{\alpha}_t}\|\Delta\epsilon\|_2,\end{aligned} \quad (13)$$

which exactly matches our proposed objective in Eq. 8.

To summarize, during training, the Smooth Diffusion objective encompasses a combination of $\mathcal{L}_{\text{base}}$ and $\mathcal{L}_{\text{reg}}$:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda\mathcal{L}_{\text{reg}}, \quad (14)$$

Figure 3. **Image interpolation comparison results.** For Smooth Diffusion and Stable Diffusion [62], real images (Image A and B) are inverted into latents using NTI [44]. We perform spherical linear interpolations between latents (also known as Stable Diffusion Walk [29]) and concatenate the resulting images as a transition sequence. VAE Inter. performs interpolations within the VAE space of Stable Diffusion. ANID [78] first adds noise to real images and subsequently denoises the interpolated noisy images using Stable Diffusion.

where $\mathcal{L}$ denotes the basic training objective of a diffusion model and represents a ratio parameter controlling the intensity of Step-wise Variation Regularization.

4. Experiments

4.1. Experimental Setup

Baselines and settings. We select the Stable Diffusion [62] as the primary baseline for all tasks. Additionally, for image interpolation, we adopt a VAE-space interpolation and ANID [78] as competitors. For image inversion, we integrate Smooth Diffusion and Stable Diffusion with DDIM inversion [71] and Null-text inversion [44]. For text-based image editing, SDEdit [42], Prompt-to-Prompt (P2P) [21], Plug-and-Play (PnP) [75], Diffusion Disentanglement (Disentangle) [82], Pix2Pix-Zero [55] and Cycle Diffusion [81] are chosen as SOTA approaches. For drag-based image editing, we compare Smooth Diffusion with Stable Diffusion within the framework of DragDiffusion [69].

Implementation details. Smooth Diffusion is trained atop pretrained Stable Diffusion-V1.5 [62], using LoRA [25] finetuning technique. The UNet of Smooth Diffusion is set as trainable with a LoRA rank of 8, while the VAE and text encoder are frozen. We leverage the LAION Aesthetics 6.5+ as the training dataset, which contains 625K image-text pairs with predicted aesthetics scores of 6.5 or higher from LAION-5B [67]. Smooth diffusion is typically trained for 30K iterations with a batch size of 96, 3 samples per GPU, a total of 4 A100 GPUs, and a gradient accumulation of 8. The AdamW [35] optimizer is adopted with a constant learning rate of $^{-}$ and a weight decay of $^{-}$. The ratio parameter in Eq. 14 is set to 1. During inference, the total number of diffusion steps is set to 50 and the classifier-free guidance [23] scale is set to 7.5.

Evaluation metrics. To evaluate the general text-to-image generation performance, we report the popular FID [22] and CLIP Score [58] on the MS-COCO validation set [38]. To assess the latent space smoothness, we propose an interpolation standard deviation (ISTD) as an evaluation metric. In specific, we randomly draw 500 text prompts from the MS-COCO validation set. For each prompt, we sample a pair of Gaussian noises and uniformly interpolate them from one to the other 9 times with mix ratios from 0.1 to 0.9. Fed into diffusion models together with a prompt, we could obtain a total of 11 generated images, 2 from the source Gaussian noises and 9 from the interpolated noises. We calculate the standard deviation of L2 distances between every two adjacent images in the pixel space. Finally, we average the standard deviations over 500 prompts as ISTD. Ideally, a zero value of ISTD indicates that consistent and uniform visual fluctuations in the pixel space for identical fixed-size changes in the latent space, resulting in a smooth latent space. For image inversion, mean square error (MSE), LPIPS [93], SSIM [79] and PSNR [27] are adopted to evaluate the image reconstruction capability.

4.2. Latent Space Interpolation

Qualitative comparison. The most straightforward way to demonstrate the smoothness of the latent space is through the observation of interpolation results between latent noises. In Fig. 3, we present interpolation compar-

isons between Smooth Diffusion and Stable Diffusion using real images. To generate these comparisons, we utilize the NTI [44] to invert a pair of real images into latent noises , sharing the same . We then perform uniform spherical linear interpolations between latent noises (also known as Stable Diffusion Walk [29]), resulting in 9 intermediate noises with mix ratios from 0.1 to 0.9. Subsequently, we concatenate the 11 images produced from these noises to create an image transition sequence in the figures.

Notably, as highlighted by the red boxes, Stable Diffusion exhibits significant visual fluctuations during the transition. In particular, the interpolated images may introduce new attributes that are unrelated to the source images, *e.g.*, the undesired grasslands in the second row of Fig. 3. In contrast, our approach, Smooth Diffusion, not only avoids introducing obvious irrelevant attributes in the interpolated images but also ensures that the visual effects change smoothly throughout the transition. Additional interpolation results can be seen in supplementary materials.

In addition to Stable Diffusion, Fig. 3 also includes two other baseline methods for comparison: 1) VAE Interpolation (VAE Inter.), which performs interpolations within the VAE space of Stable Diffusion. However, the results closely resemble pixel-space interpolations, with significant degradation of visual details, particularly in the highlighted red box area. 2) ANID [78], which first adds noise to real images and subsequently denoises the interpolated noisy images using Stable Diffusion. In Fig. 3, ANID with a 50-step scheduler exhibits highly blurred interpolation results. When ANID operates with a default 200-step scheduler, the blurring can be alleviated, but the quality of the interpolated images remains far from satisfactory.

Method	ISTD ( )	FID ( )	CLIP Score ( )
Stable Diffusion	38.63	12.70	31.46
Smooth Diffusion	**16.54**	**12.10**	**31.54**

Table 1. **Quantitative evaluations of image interpolation and text-to-image generation.** We evaluate Smooth Diffusion and Stable Diffusion [62] with ISTD, FID [22] and CLIP Score [58]. The better results are in bold.

Quantitative comparison. The goal of Smooth Diffusion is to enhance the latent space smoothness without image generation performance degradation compared to Stable Diffusion. In pursuit of this goal, we employ the ISTD introduced in Sec. 4.1 to evaluate the latent space smoothness. Additionally, we utilize FID [22] and CLIP Score [58] to assess generators' overall performance. The results presented in Tab. 1 demonstrate that Smooth Diffusion significantly outperforms Stable Diffusion in terms of ISTD, indicating a substantial improvement in the latent space smoothness. Furthermore, Smooth Diffusion exhibits superior performance in both FID and CLIP Score, suggesting that the enhancement of latent space smoothness and the overall image generation quality are not mutually exclusive but complement each other when the regularization term is applied with a suitable strength ratio.

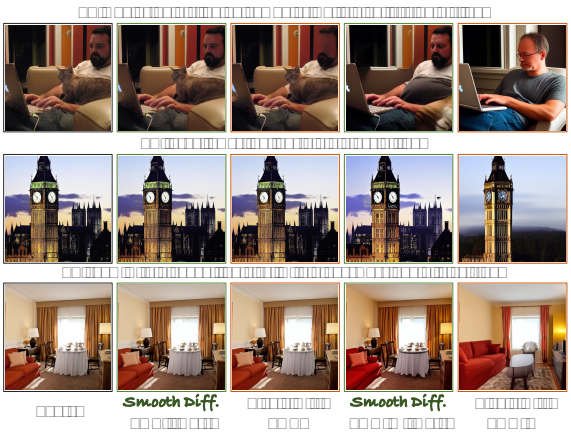


Figure 4. **Image reconstruction comparison results.** We integrate Smooth Diffusion and Stable Diffusion [62] with NTI [44] (column 2 & 3) and DDIM inversion [71] (column 4 & 5).

Method	MSE ( )	LPIPS ( )	SSIM ( )	PSNR ( )
Stable Diff. + DDIM	0.1756	0.5385	0.2662	13.97
Smooth Diff. + DDIM	0.1086	0.4326	0.3418	16.17
Stable Diff. + NTI	0.0156	0.1656	0.6068	25.63
Smooth Diff. + NTI	**0.0153**	**0.1635**	**0.6102**	**25.74**
VAE Reconstruction	0.0148	0.1590	0.6136	25.98

Table 2. **Quantitative evaluations of image reconstruction.** We integrate Stable Diffusion and Smooth Diffusion [62] with DDIM inversion [71] (row 2 & 3) and NTI [44] (row 4 & 5). MSE, LPIPS [93], SSIM [79] and PSNR [27] are evaluated. VAE Reconstruction results are provided as the optimal values.

4.3. Image Inversion and Reconstruction

Previous research [32] in the realm of GANs discovered that a smoother latent space has a positive impact on the accuracy of image inversion and reconstruction. We empirically validate this finding within the context of diffusion models. In specific, two representative inversion techniques, DDIM inversion [71] and Null-text inversion (NTI) [44] are adopted and integrated with Smooth Diffusion and Stable Diffusion separately. We both qualitatively and quantitatively compare the image inversion and reconstruction performance of these integrated models using 500 randomly sampled images from the MS-COCO validation set [38].

As illustrated in the two rightmost columns of Fig. 4, when employing a straightforward DDIM inversion, Smooth Diffusion outperforms Stable Diffusion by a considerable margin in terms of reconstruction quality. This improvement is evident in various aspects, such as an accurate generation of character identities, a faithful recreation of the city view behind the tower, and a correct reproduc-

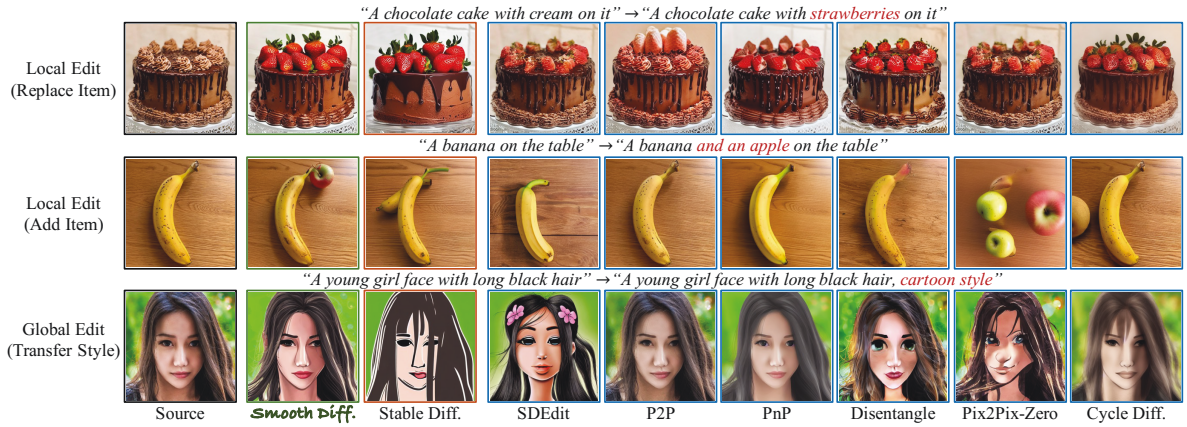


Figure 5. **Text-based image editing comparison results.** We compare Smooth Diffusion and Stable Diffusion [62] (column 2 & 3), considering both local and global edits through the straightforward pipeline described in Sec. 4.4. Additionally, we present results from SOTA approaches, including SDEdit [42], P2P [21], PnP [75], Disentangle [82], Pix2Pix-Zero [55], and Cycle Diffusion [81], as references.

tion of room layouts. This phenomenon underscores the fact that the latent space of Smooth Diffusion is more tolerant of the errors introduced by the local linear approximation in DDIM inversion. Consequently, the reconstruction results produced by Smooth Diffusion manage to retain the contents of the source images to a greater extent. On the other hand, when the optimization-based NTI technique is employed, the disparity between Smooth Diffusion and Stable Diffusion is not as pronounced. Nonetheless, there are still instances where Stable Diffusion exhibits subpar results, such as the ruined man's face in Fig. 4.

To quantify the image reconstruction performance, MSE, LPIPS [93], SSIM [79] and PSNR [27] are reported in Tab. 2. Notably, the reconstruction error encompasses two components: 1) the error from different inversion methods and U-Net parameters and 2) the error from the shared pretrained VAE [36]. Hence, we included the VAE reconstruction errors as optimal values for our method. The results exhibit a consistent outperformance of Smooth Diffusion over Stable Diffusion across all metrics, whether using DDIM inversion or NTI. Moreover, "Smooth Diffusion + NTI" performs results close to VAE reconstruction, indicating its superiority attributed to a smoother latent space.

4.4. Image Editing

The superiority of Smooth Diffusion in image inversion and reconstruction has motivated us to explore its potential for enhancing image editing tasks. In this section, we delve into two typical image editing scenarios: text-based image editing and drag-based image editing.

Text-based image editing. There have been numerous methods [21, 42, 55, 75, 81, 82] proposed in the literature, each with its own unique designs aimed at achieving the SOTA performance. In contrast, we adopt a simpler pipeline akin to the image inversion and reconstruction process discussed in Sec. 4.3. The key distinction lies in our approach to modify the text prompt during the later time steps of the reconstruction process. In specific, the original $\epsilon_\theta(\boldsymbol{x}_t, t, \mathcal{C}, \varnothing_t)$ in Eq. (5) during NTI reconstruction (diffusion sampling) process is replaced with:

$$\epsilon_\theta(\boldsymbol{x}_t, t, \mathcal{C}, \varnothing_t) = \begin{cases} \epsilon_\theta(\boldsymbol{x}_t, t, \mathcal{C}_{\text{src}}, \varnothing_t), & t > T \times r, \\ \epsilon_\theta(\boldsymbol{x}_t, t, \mathcal{C}_{\text{trg}}, \varnothing_t), & t \leq T \times r, \end{cases} \tag{15}$$

where $\mathcal{C}_{\text{src}}$ represents the source text prompt for inversion, while $\mathcal{C}_{\text{trg}}$ corresponds to the target text prompt for editing. The parameter r serves as a threshold, determining when to switch from $\mathcal{C}_{\text{src}}$ to $\mathcal{C}_{\text{trg}}$. In practice, r is typically chosen within $\{0.6, 0.7, 0.8, 0.9\}$, with the exact value depending on the specific input images and target visual effects.

Through this straightforward pipeline, we conducted a comparative analysis of the editing performance between Smooth Diffusion and Stable Diffusion, as presented in the three left-most columns of Fig. 5. We also included editing results obtained from SOTA methods as references. Our evaluation encompasses both local and global editing tasks. The local editing tasks involve replacing items (*e.g.*, changing "cream" to "strawberries") and adding items (*e.g.*, "apple"). On the other hand, the global editing tasks pertain to global style transfer, such as transforming an image into a "cartoon style". It is evident that while Stable Diffusion excels in achieving precise image reconstruction with NTI, as discussed in Sec. 4.3, even minor modifications to the text prompt can significantly impact the content of the generated images. For instance, it can affect elements like the style of the cake, the shape of the banana, and the haircut of the girl. In contrast, Smooth Diffusion not only accurately generates edited images in accordance with the target text prompts but also effectively preserves the unedited contents. Furthermore, when compared to SOTA methods, even with this straightforward pipeline, Smooth Diffusion consistently delivers competitive results across all cases.

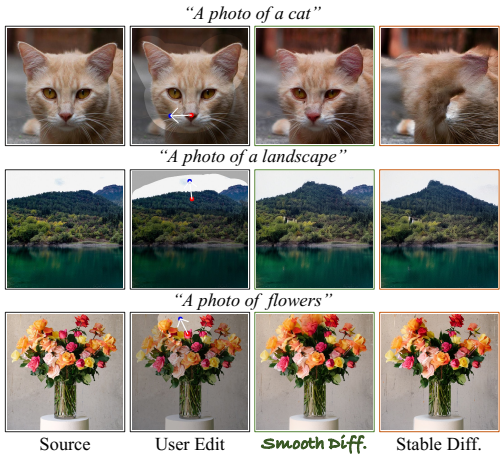


Figure 6. **Drag-based image editing comparison results.** We implement Smooth Diffusion and Stable Diffusion [62] within the framework of DragDiffusion [69], respectively.

Drag-based image editing. As an emerging research avenue in the community, drag-based image editing [45, 53, 69] has garnered considerable attention recently. DragDiffusion [69] first introduces a framework for drag-based image editing employing Stable Diffusion. In the task 3 of Fig. 1 and Fig. 6, we showcase that by integrating Smooth Diffusion into the DragDiffusion framework, some previously unsuccessful editing operations with Stable Diffusion can be enabled. As illustrated, Smooth Diffusion achieves operations such as making the tree grow taller without damaging existing branches (Fig. 1), rotating the cat head, creating a new mountain top without destroying the original one, and letting new flowers grow in the vase (Fig. 6). These operations, however, fail with Stable Diffusion, indicating the non-smoothness of its latent space.

4.5. Ablation Studies

Regularization ratio. In Tab. 3, we examine the impact of different strength ratios in Eq. (14). This ratio adjusts the intensity of the step-wise variation regularization. Specifically, when a weaker regularization is applied (*e.g.*,), we observe a slight improvement in the CLIP Score. However, there is a significant increase in ISTD, indicating a notable degradation in latent space smoothness. In contrast, employing a stronger regularization (*e.g.*,) leads to a smoother latent space, as demonstrated by the decrease in ISTD. However, in this case, we observe an unexpected increase in FID, indicating a notable decline in the quality of generated images. Therefore, selecting an appropriate trade-off value for becomes crucial based on the specific experimental settings. In our default setting, we find that serves as a suitable value.

LoRA rank. In Tab. 4, we examine the impact of different ranks of the LoRA component utilized in our Smooth diffusion. We discover that LoRA ranks within the range of [4,16] are all suitable values for our default setting. We select a default rank of 8 because of its lowest ISTD among the first three rows in Tab. 4. Furthermore, we train a fully finetuned model, referred to as "full," which showcases a further decrease in ISTD. However, this comes at the expense of significantly degrading the quality of the generated images, as indicated by an increased FID and decreased CLIP Score. This decline in performance underscores the vulnerability of fully fine-tuned models to collapse within our default setting, emphasizing the need for additional meticulous design considerations.

Ratio	ISTD ( )	FID ( )	CLIP Score ( )
0.1	24.23	<u>12.15</u>	**31.56**
1 (default)	<u>16.54</u>	**12.11**	<u>31.49</u>
10	**11.51**	17.44	31.41

Table 3. **Ablation results of different regularization ratios.** The best results are in bold, and the second-best results are underlined.

Rank	ISTD ( )	FID ( )	CLIP Score ( )
4	16.76	12.36	31.49
8 (default)	<u>16.54</u>	<u>12.11</u>	<u>31.54</u>
16	16.65	**11.49**	**31.61**
full	**11.52**	27.27	28.86

Table 4. **Ablation results of different LoRA ranks.** The best results are in bold, and the second-best results are underlined.

5. Conclusion

In this article, we explored Smooth Diffusion, an innovative diffusion model that enhances latent space smoothness for generation. Smooth Diffusion adopts the novel Step-wise Variation Regularization, which successfully maintains variation between arbitrary input latent and generated images at a more bounded range. Smooth Diffusion was trained on top of the prevailing text-to-image model, from which we carried out extensive research, including but not limited to interpolation, inversion, and editing, all of which had shown competitive performance. Through qualitative and quantitative measurements, we demonstrated that Smooth Diffusion managed to make a smoother latent space without compromising the output quality. We believe that Smooth Diffusion will become a valuable solution for other challenging tasks, such as video generation, in the future.

Acknowledgements

This work is supported in part by the ML Center at Georgia Tech, NSF CAREER Award #2239840, and the National AI Institute for Exceptional Education (Award #2229873) by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education. Huang is in part supported by the National Key R&D Program (#2021ZD0140407) and NSFC (62321005 & 42327901).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN++: How to edit the embedded images? In *CVPR*, 2020. 3

[2] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *ICCV*, 2021. 3

[3] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv:1701.04862*, 2017. 3

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 3

[5] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *ICLR*, 2022. 2

[6] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *ICML*, 2023. 2

[7] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *ICLR*, 2018. 3

[8] Yoshua Bengio, Eric Laufer, Guillaume Alain, and Jason Yosinski. Deep generative stochastic networks trainable by backprop. In *ICML*, 2014. 2

[9] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv:1809.11096*, 2018. 3

[10] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, 2021. 3

[11] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 2, 3

[13] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. *arXiv:2305.04441*, 2023. 2

[14] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv:2303.17546*, 2023. 2

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2, 3

[16] 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 *CVPR*, 2022. 2

[17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *NeurIPS*, 30, 2017. 3

[18] Jiayi Guo, Chaoqun Du, Jiangshan Wang, Huijuan Huang, Pengfei Wan, and Gao Huang. Assessing a single image in reference-guided image synthesis. In *AAAI*, 2022. 3

[19] Jiayi Guo, Hayk Manukyan, Chenyu Yang, Chaofei Wang, Levon Khachatryan, Shant Navasardyan, Shiji Song, Humphrey Shi, and Gao Huang. Faceclip: Facial image-to-video translation via a brief text description. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[20] Jiayi Guo, Chaofei Wang, You Wu, Eric Zhang, Kai Wang, Xingqian Xu, Humphrey Shi, Gao Huang, and Shiji Song. Zero-shot generative model adaptation via image-specific prompt learning. In *CVPR*, 2023. 2

[21] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv:2208.01626*, 2022. 2, 5, 7

[22] 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. In *NeurIPS*, 2017. 5, 6

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshops*, 2021. 3, 5

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 2, 3, 5

[26] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv:2302.09778*, 2023. 2

[27] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 2008. 5, 6, 7

[28] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, 2023. 2

[29] Andrej Karpathy. Stable Diffusion Walk. https://gist.github.com/karpathy/00103b0037c5aaea32fe1da1af553355, 2022. 2, 5, 6

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[31] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020. 3

[32] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2, 3, 4, 6

[33] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 2, 3

[34] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 2

[35] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *ICLR*, 2015. 5

[36] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2015. 2, 7

[37] Chieh Hubert Lin, Hsin-Ying Lee, Yen-Chi Cheng, Sergey Tulyakov, and Ming-Hsuan Yang. Infinitygan: Towards infinite-pixel image synthesis. In *ICLR*, 2021. 3

[38] 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 *ECCV*, 2014. 5, 6

[39] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022. 2

[40] Haoming Lu, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style. In *CVPR*, 2023. 2

[41] Weijian Mai and Zhijun Zhang. Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. *arXiv:2308.07428*, 2023. 2

[42] 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 *ICLR*, 2022. 2, 5, 7

[43] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *CVPR*, 2023. 2

[44] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 2, 3, 5, 6

[45] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv:2307.02421*, 2023. 2, 8

[46] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv:2302.08453*, 2023. 2

[47] Zanlin Ni, Yulin Wang, Renping Zhou, Jiayi Guo, Jinyi Hu, Zhiyuan Liu, Shiji Song, Yuan Yao, and Gao Huang. Revisiting non-autoregressive transformers for efficient image synthesis. In *CVPR*, 2024. 2

[48] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2

[49] 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 *ICML*, 2022. 2

[50] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Face identity disentanglement via latent space mapping. *TOG*, 2020. 3

[51] Augustus Odena, Jacob Buckman, Catherine Olsson, Tom Brown, Christopher Olah, Colin Raffel, and Ian Goodfellow. Is generator conditioning causally related to gan performance? In *ICML*, 2018. 3, 4

[52] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *TPAMI*, 2021. 3

[53] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your GAN: Interactive point-based manipulation on the generative image manifold. In *SIGGRAPH*, 2023. 3, 8

[54] Zhihong Pan, Riccardo Gherardi, Xiufeng Xie, and Stephen Huang. Effective real image editing with accelerated iterative diffusion inversion. In *ICCV*, 2023. 2

[55] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *SIGGRAPH*, 2023. 2, 5, 7

[56] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-driven manipulation of StyleGAN imagery. In *ICCV*, 2021. 3

[57] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv:2305.11147*, 2023. 2

[58] 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 *ICML*, 2021. 2, 5, 6

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

[60] 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 *CVPR*, 2021. 3

[61] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 3

[62] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 5, 6, 7, 8

[63] 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. In *CVPR*, 2023. 2

[64] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[65] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 2

[66] Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *ICML*, 2015. 2

[67] 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. In *NeurIPS*, 2022. 5

[68] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. InterfaceGAN: Interpreting the disentangled face representation learned by GANs. *TPAMI*, 2020. 2, 3

[69] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. DragDiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv:2306.14435*, 2023. 2, 5, 8

[70] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[71] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 2, 3, 5, 6

[72] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023. 2

[73] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *arXiv:2305.11846*, 2023. 2

[74] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *TOG*, 2021. 3

[75] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023. 2, 5, 7

[76] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. 2

[77] Bram Wallace, Akash Gokul, and Nikhil Naik. EDICT: Exact diffusion inversion via coupled transformations. In *CVPR*, 2023. 2

[78] Clinton Wang and Polina Golland. Interpolating between images with diffusion models. In *ICML Workshops*, 2023. 2, 5, 6

[79] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 5, 6, 7

[80] Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. *arXiv:2305.01115*, 2023. 2

[81] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *ICCV*, 2023. 2, 5, 7

[82] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *CVPR*, 2023. 2, 5, 7

[83] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Wensen Feng. Controllable continuous gaze redirection. In *ACM MM*, 2020. 3

[84] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *TPAMI*, 2022. 3

[85] Dejia Xu, Xingqian Xu, Wenyan Cong, Humphrey Shi, and Zhangyang Wang. Reference-based painterly inpainting via diffusion: Crossing the wild reference domain gap. *arXiv:2307.10584*, 2023. 2

[86] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. *arXiv:2305.16223*, 2023. 2

[87] Xingqian Xu, Shant Navasardyan, Vahram Tadevosyan, Andranik Sargsyan, Yadong Mu, and Humphrey Shi. Image completion with heterogeneously filtered spectral hints. In *WACV*, 2023. 3

[88] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *ICCV*, 2023. 2

[89] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv:2308.06721*, 2023. 2

[90] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023.

[91] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2

[92] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022. 2

[93] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 6, 7

[94] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, I Eric, Chao Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *ICLR*, 2020. 3

[95] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-ControlNet: All-in-one control to text-to-image diffusion models. In *NeurIPS*, 2023. 2

[96] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *NeurIPS*, 2023. 2

[97] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *ECCV*, 2020. 3