SINE: SINgle Image Editing with Text-to-Image Diffusion Models

Zhixing Zhang[1] Ligong Han[1] Arnab Ghosh[2] Dimitris Metaxas[1] Jian Ren[2]
[1]Rutgers University [2]Snap Inc.

Abstract

Recent works on diffusion models have demonstrated a strong capability for conditioning image generation, e.g., text-guided image synthesis. Such success inspires many efforts trying to use large-scale pre-trained diffusion models for tackling a challenging problem–real image editing. Works conducted in this area learn a unique textual token corresponding to several images containing the same object. However, under many circumstances, only one image is available, such as the painting of the Girl with a Pearl Earring. Using existing works on fine-tuning the pre-trained diffusion models with a single image causes severe overfitting issues. The information leakage from the pre-trained diffusion models makes editing can not keep the same content as the given image while creating new features depicted by the language guidance. This work aims to address the problem of single-image editing. We propose a novel model-based guidance built upon the classifier-free guidance so that the knowledge from the model trained on a single image can be distilled into the pre-trained diffusion model, enabling content creation even with one given image. Additionally, we propose a patch-based fine-tuning that can effectively help the model generate images of arbitrary resolution. We provide extensive experiments to validate the design choices of our approach and show promising editing capabilities, including changing style, content addition, and object manipulation. Our code is made publicly available here.

1. Introduction

Automatic real image editing is an exciting direction, enabling content generation and creation with minimal effort. Although many works have been conducted in this area, achieving high-fidelity semantic manipulation on an image is still a challenging problem for the generative models, considering the target image might be out of the training data distribution [4, 11, 25, 26, 32, 58, 59]. The recently introduced large-scale text-to-image models, *e.g.*, DALL·E 2 [37], Imagen [42], Parti [55], and StableDiffusion [39], can perform high-quality and diverse image generation with natural language guidance. The success of these works has inspired many subsequent efforts to leverage the pre-trained large-scale models for real image editing [10, 15, 20, 40]. They show that, with properly designed prompts and a limited number of fine-tuning steps, the text-to-image models can manipulate a given subject with text guidance.

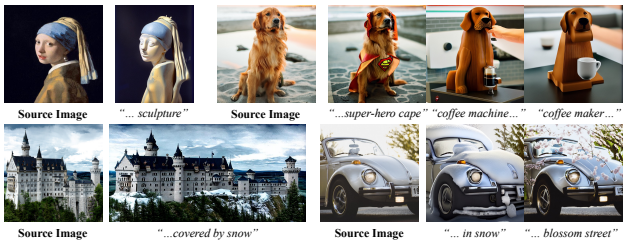


Figure 1. With only *one* real image, *i.e.*, Source Image, our method is able to manipulate and generate the content in various ways, such as changing style, adding context, modifying the object, and enlarging the resolution, through guidance from the text prompt.

On the downside, the recent text-guided editing works that build upon the diffusion models suffer several limitations. First, the fine-tuning process might lead to the pre-trained large-scale model overfit on the real image, which degrades the synthesized images' quality when editing. To tackle these issues, methods like using multiple images with the same content and applying regularization terms on the same object have been introduced [4, 40]. However, querying multiple images with identical content or object might not be an available choice; for instance, there is only one painting for *Girl with a Pearl Earring*. Directly editing the single image brings information leakage from the pre-trained large-scale models, generating images with different content (examples in Fig. 5); therefore, the application scenarios of these methods are greatly constrained. Second, these works lack a reasonable understanding of the object geometry for the edited image. Thus, generating images with different spatial size as the training data cause undesired artifacts, *e.g.*, repeated objects, and incorrectly modified geometry (examples in Fig. 5). Such drawbacks restrict applying these methods for generating images with an arbitrary resolution, *e.g.*, synthesizing high-resolution images from a single photo of a castle (as in Fig. 4), again limiting

the usage of these methods.

In this work, we present SINE, a framework utilizing pre-trained text-to-image diffusion models for **SIN**gle image **E**diting and content manipulation. We build our approach based upon existing text-guided image generation approaches [10, 40] and propose the following novel techniques to solve overfitting issues on content and geometry, and language drift [28, 31]:

- First, by appropriately modifying the classifier-free guidance [22], we introduce model-based classifier-free guidance that utilizes the *diffusion model* to provide the score guidance for content and structure. Taking advantage of the step-by-step sampling process used in diffusion models, we use the model fine-tuned on a single image to plant a content "seed" at the early stage of the denoising process and allow the pre-trained large-scale text-to-image model to edit creatively conditioned with the language guidance at a later stage.
- Second, to decouple the correlation between pixel position and content, we propose a patch-based fine-tuning strategy, enabling generation on arbitrary resolution.

With a text descriptor describing the content that is aimed to be manipulated and language guidance depicting the desired output, our approach can edit the *single* unique image to the *targeted domain* with details preserved in *arbitrary* resolution. The output image keeps the structure and background intact while having features well-aligned with the target language guidance. As shown in Fig. 1, trained on a painting *Girl with a Pearl Earring* with resolution as 512×512, we can sample an image of a sculpture of the girl at the resolution of 640×512 with the identity features preserved. Moreover, our method can successfully handle various edits such as style transfer, content addition, and object manipulation (more examples in Fig. 3). We hope our method can further boost creative content creation by opening the door to editing arbitrary images.

2. Related Work

Text-guided image synthesis has drawn considerable attention in the generative model context [1, 2, 6, 12, 14, 16–19, 34, 35, 38, 50, 53–56, 60]. The recent development of diffusion models [21, 46, 47, 49] introduced new solutions to this problem and produced impressive results [33, 37, 39, 42]. With the significant improvement of these models, rather than training a large-scale text-to-image model from scratch, a leading line of works focuses on taking advantage of the existing pre-trained model and manipulating images according to given natural language guidance [3, 10, 20, 24, 29, 40]. In these works, studies explore text-based interfaces for image editing [1, 3, 35], style transfer [27, 30], and generator domain adaption [11, 25, 48].

The development of the diffusion model provides a giant and flexible design space for this task. Many works utilize pre-trained diffusion models as generative priors and are training-free. ILVR [7] guides the denoising process by replacing the low-frequency part of the sample with that of the target reference image. SDEdit [32] applies the diffusion process first on an image or a user-created semantic map and then conducts the denoising procedure conditioned with the desired output. Blended diffusion [3] performs language-guided inpainting with a given mask.

Another line of research showed great potential and semantic editing ability of fine-tuning. DiffusionCLIP [25] leverages the CLIP [36] model to provide gradients for image manipulation and deliver impressive results on style transfer. Textual-Inversion [10] and DreamBooth [40] fine-tune the text embedding or the full diffusion model using a few personalized images (typically $3 \sim 5$) to synthesize images of the same object in a novel context. These methods, however, either drastically change the layout of the original image when dealing with a single image or can not fully leverage the generalization ability of the pre-trained model for editing due to overfitting or language drift. Notably, Prompt-to-Prompt [20] controls the editing of synthesized images by manipulating the cross-attention maps; however, its editing ability is limited when applied to real images. Imagic [24], while delivering realistic results, requires fine-tuning process on each editing prompt for each image, leading to longer inference time.

This work introduces a solution to achieve image fidelity and text alignment simultaneously. Our method can perform high-quality semantic editing globally and locally on one single image. On the other hand, previous works lack an understanding of the object geometry of the edited image. When editing the image at an arbitrary resolution, the artifacts in the results will be obvious. Prior works have investigated generating images at arbitrary resolution using positional encoding as inductive bias [41, 52, 52] so that the correlation between content and position can be eliminated. Anyres-GAN [5] adopt a patch training mechanism to leverage high-resolution data to help the generation of images in the low-resolution domain. We propose a patch-based fine-tuning method to achieve arbitrary resolution editing.

3. Methods

For *one* arbitrary in-the-wild image, our goal is to edit the image via language while preserving the maximal amount of details from the original image. To do so, we leverage the generalization ability of pre-trained large-scale text-to-image models [39]. An intuitive approach is to fine-tune the diffusion models with the single image and text description, similar to DreamBooth [40]. Ideally, it should provide a model that can reconstruct the input image using the given text descriptor and synthesize new images when

given other language guidance. Unfortunately, we find the model can easily overfit the single trained image and its corresponding text description. Thus, although the fine-tuned model can still reconstruct the input image perfectly, it can no longer synthesize diverse images according to the given language guidance (as shown in Fig. 5). Moreover, it struggles to generate arbitrary resolution images due to the lack of positional information (as in Fig. 4).

To solve the above issues, we propose a test-time model-based classifier-free guidance and a patch-based fine-tuning technique. An overview of our method is illustrated in Fig. 2. In the following sections, we review the backbone model used in our approach (Sec. 3.1). Then, we describe how to overcome the overfitting problem with model-based guidance (Sec. 3.2). Lastly, we present how to address the problem of limited resolution generation (Sec. 3.3).

3.1. Language-Guided Diffusion Models

We use the latent diffusion models (LDMs) [39] trained on a large-scale dataset as our base model and implement the proposed approaches by fine-tuning the pre-trained model. LDMs is a class of Denoising Diffusion Probabilistic Models (DDPMs) [21] that contains an auto-encoder trained on images, and a diffusion model learned on the latent space constructed by the auto-encoder. The encoder $\mathcal{E}$ encodes a given image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ to a latent representation $\mathbf{z}$, such that $\mathbf{z} = \mathcal{E}(\mathcal{I})$. The decoder $\mathcal{D}$ reconstructs the estimated image $\tilde{\mathcal{I}}$ from the latent, such that $\tilde{\mathcal{I}} = \mathcal{D}(\mathbf{z})$ and $\tilde{\mathcal{I}} \approx \mathcal{I}$. The diffusion model is trained to produce latent codes within the pre-trained latent space. The most intriguing property of LDMs is that the diffusion model can be conditioned on class labels, images, and text prompt. The conditional LDM is learned as follows:

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(\mathcal{I}),y,\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\mathbf{z}_t, t, \boldsymbol{\tau}_\theta(y)\right)\|_2^2\right], \quad (1)$$

where t is the time step, $\mathbf{z}_t$ is the latent noised to time t, $\boldsymbol{\epsilon}$ is the unscaled noise sample, $\boldsymbol{\epsilon}_\theta$ is the denoising model, y is the conditioning input, and $\boldsymbol{\tau}_\theta$ maps y to a conditioning vector. During training time, $\boldsymbol{\epsilon}_\theta$ and $\boldsymbol{\tau}_\theta$ are jointly optimized. A random noise tensor is sampled and denoised at inference time based on the conditioning input, *e.g.*, text prompt, to produce a new latent. Inspired by DreamBooth [40], we construct the text prompt for fine-tuning a single image as "a photo/painting of a [$*$] [class noun]", where "[$*$]" is a unique identifier and "[class noun]" is a coarse class descriptor (*e.g.*, "castle", "lake", "car", *etc.*).

3.2. Model-Based Classifier-Free Guidance

With the above-presented LDMs, we introduce our approach, inspired by classifier-free guidance, to overcome overfitting when fine-tuning LDMs with one image.

Classifier-free guidance [22] is a technique widely adopted by prior text-to-image diffusion models [39, 42]. A single

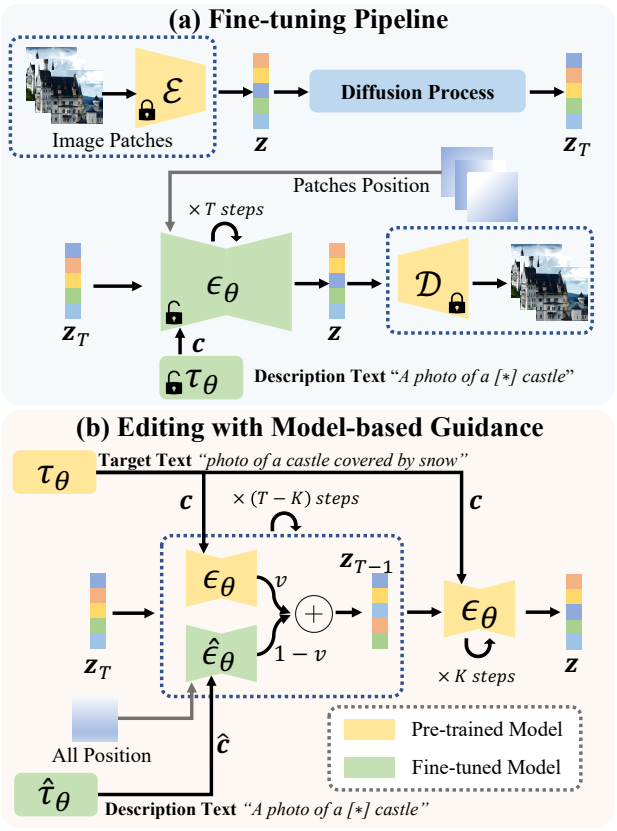


Figure 2. **Overview of our method.** (a) Given a source image, we first randomly crop it into patches and get the corresponding latent code $\mathbf{z}$ with the pre-trained encoder. At fine-tune time, the denoising model, $\boldsymbol{\epsilon}_\theta$, takes three inputs: noisy latent $\mathbf{z}_T$, language condition $\mathbf{c}$, and positional embedding for the area where the noisy latent is obtained. (b) During sampling, we give additional language guidance about the target domain to edit the image. Also, we sample a noisy latent code $\mathbf{z}_T$ with the dimension corresponding to the desired output resolution. Language conditioning for $\boldsymbol{\epsilon}_\theta$ and $\mathbf{c}$ are given by pre-trained language encoder $\boldsymbol{\tau}_\theta$ with the target language guidance. While for the fine-tuned diffusion model, $\hat{\boldsymbol{\epsilon}}_\theta$, in addition to the language conditioning $\hat{\mathbf{c}}$, we also input the positional embedding for the whole image. We employ a linear combination between the score calculated by each model for the first K steps and inference only on pre-trained $\boldsymbol{\epsilon}_\theta$ after.

diffusion model is trained using conditional and unconditional objectives by randomly dropping the condition during training. When sampling, a linear combination of the conditional and unconditional score estimation is used:

$$\tilde{\boldsymbol{\epsilon}}_\theta\left(\mathbf{z}_t, \mathbf{c}\right) = w\boldsymbol{\epsilon}_\theta\left(\mathbf{z}_t, \mathbf{c}\right) + (1-w)\boldsymbol{\epsilon}_\theta\left(\mathbf{z}_t\right), \quad (2)$$

where $\boldsymbol{\epsilon}_\theta\left(\mathbf{z}_t, \mathbf{c}\right)$ and $\boldsymbol{\epsilon}_\theta\left(\mathbf{z}_t\right)$ are the conditional and unconditional $\boldsymbol{\epsilon}$-predictions, $\mathbf{c}$ is the conditioning vector generated by τ_θ, and w is the weight for the guidance. The predication is performed using the Tweedie's formula [8], namely, $\left(\mathbf{z}_t - \sigma_t\tilde{\boldsymbol{\epsilon}}_\theta\right)/\alpha_t$, where α_t and σ_t are functions of

t that affect the sampling quality.

Since we only have one image as the training data, *e.g.*, painting of *Mona Lisa*, and one corresponding text descriptor of that image, the diffusion model suffers from overfitting, and severe language drifts after fine-tuning [40]. As a result, the fine-tuned model fails to synthesize images containing features from other language guidance. The overfitting issue might be due to only one repeated prompt used during fine-tuning, making other text prompts no longer accurate enough to control editing (see examples in Fig. 6).

Model-based classifier-free guidance. Existing "personalized" text-guided real image editing works only use *one* fine-tuned model for image generation and editing [10, 20, 40], ignoring the capacity of pre-trained large-scale text-to-image models. Instead, to alleviate the overfitting of the fine-tuned model, we leverage the *pre-trained text-to-image model for image generation* with the provided language guidance and use the fine-tuned model to provide content features in a fashion of combining scores from the two models, similar to classifier-free guidance.

Specifically, let $\hat{\boldsymbol{\epsilon}}_\theta$ denote the fine-tuned denoising model, and $\boldsymbol{\epsilon}_\theta$ denote the pre-trained text-to-image model. During sampling, at specified steps, we use our fine-tuned model to guide the pre-trained one by using a linear combination of the scores from each model. Thus, the score estimation in Eqn. 2 becomes:

$$\begin{aligned}\tilde{\boldsymbol{\epsilon}}_\theta\left(\mathbf{z}_t, \mathbf{c}\right) =& w\left(v\boldsymbol{\epsilon}_\theta\left(\mathbf{z}_t, \mathbf{c}\right) + (1-v)\hat{\boldsymbol{\epsilon}}_\theta\left(\mathbf{z}_t, \hat{\mathbf{c}}\right)\right) \\ &+ (1-w)\boldsymbol{\epsilon}_\theta\left(\mathbf{z}_t\right),\end{aligned} \tag{3}$$

where v stands for the model guidance weight, $\hat{\mathbf{c}}$ is the language guidance token obtained from the fine-tuned diffusion model with the text prompt used during fine-tuning, and $\mathbf{c}$ is the target language conditioning obtained from the target prompt.

To prevent artifacts from the over-fitted model and maintain the fidelity of the generated image, we propose to sample using Eqn. 3 with $t > K$ and sample using Eqn. 2 for $t \leq K$. From K to 0, the denoising process only depends on the pre-trained model. Following this approach, we can fully leverage the generalization ability of the pre-trained model (examples in Fig. 6). Also note that this method could be generalized to include multiple prompts or even multiple modalities.

3.3. Patch-Based Fine-Tuning

With model-based classifier-free guidance, we are able to edit and manipulate a single image with given language guidance. Here, we further show how to improve the fine-tuning process for a single training image so that the fine-tuned model can better understand the content and geometry of the image. Thus, it can provide better content guidance for the large-scale text-to-image model during the sampling time and unleash the potential for generating arbitrary-resolution images [9].

Limited-resolution generation. We first review the limitations of the current fine-tuning process. Given an input image $\mathcal{I}$ with resolution as $H \times W$, we can obtain a downsampled latent code $\mathbf{z}$ from the pre-trained encoder. Since the text-to-image diffusion model is pre-trained at a fixed resolution, *i.e.*, $p \times p$, we need to resize the input image to a corresponding resolution $sp \times sp$, where s represents the scaling factor of the encoder, to match the resolution for reducing the fine-tuning cost. In essence, prior knowledge of the correlation between the position and content information is learned by the diffusion model. Thus, when sampling from a higher-resolution noise tensor, the generated latent code leads to artifacts like duplicates or position shifting (visual examples in Fig. 5). To tackle such drawbacks, we propose a simple yet effective fine-tuning method.

Patch-based fine-tuning. Inspired by Chai *et al.* [5], we treat our single training image as a function on coordinate for each pixel, bounded in $[0, H] \times [0, W]$. The diffusion model still generates latent code at the fixed resolution $p \times p$, but each latent code corresponds to a sub-area in the image. We denote the sub-area as $\mathbf{v} = [h_1, w_1, h_2, w_2]$, where $(h_1, w_1) \in [0, H] \times [0, W]$ and $(h_2, w_2) \in (h_1, H] \times (w_1, W]$ indicate the top-left and bottom-right coordinates of the area, respectively. During fine-tuning, we sample patches from the image with different $\mathbf{v}$ and resize the patches to resolution sp. We denote the resulted patch as $\mathcal{I}(F(\mathbf{v})) \in \mathbb{R}^{sp \times sp \times 3}$, where F is the normalization and Fourier embedding [23] of the specific area. The encoded latent code of the patch is $\mathbf{z}_\mathbf{v} = \mathcal{E}(\mathcal{I}(F(\mathbf{v})))$. Our model uses the normalized Fourier embedding as an input to make the model learn the position-content correlation. Formally, our diffusion model is defined as $\hat{\boldsymbol{\epsilon}}_\theta\left(\mathbf{z}_t, t, \boldsymbol{\tau}_\theta(y), F(\mathbf{v})\right)$.

After fine-tuning, the model can generate latent code at different resolutions by giving the positional information directly to the model. The arbitrary resolution image editing is conducted by feeding two inputs to the model: the positional embedding of the whole image; and a randomly sampled noisy latent with the dimension corresponding to the resolution we want. When sampling in an arbitrary resolution, the model can still keep the structure of the original image intact (examples in Fig. 4). It is worth noting that with or without the correct position encoding, our framework still naturally permits retargeting, *i.e.*, maintaining the aspect ratio of salient objects, like SinGAN [44], InGAN [45], and Drop-the-GAN [13].

4. Experiments

Implementation Details While our method can be generally applied to different frameworks, we implement it based on the recently released text-to-image LDM, Stable Diffusion [39]. The pre-trained model was trained on 512×512

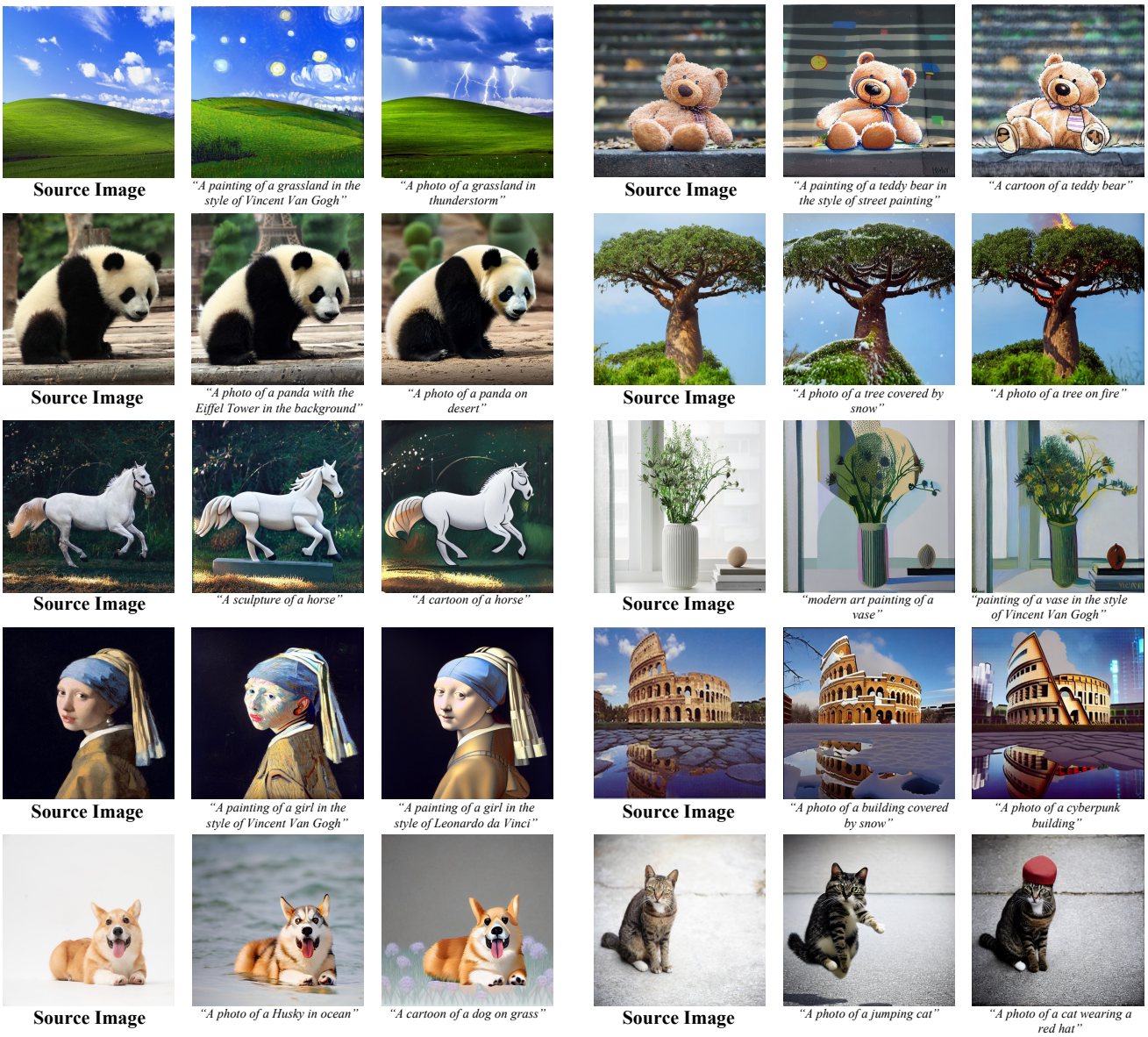


Figure 3. **Editing on single source image from various domains.** We employ our method on various images and edit them with two target prompts at 512×512 resolution. We show the wide range of edits our approach can be used, including but not limited to style transfer, content add-on, posture change, breed change, *etc*.

images from LAION dataset [43]. The spatial size of the latent code from the pre-trained model is 64×64.

For patch-based fine-tuning, we randomly crop images to patches with height and width uniformly in the range of $[0.1H, H] \times [0.1W, W]$ and resize them to 512×512. Experiments are conducted using $1 \times$ RTX 8000 GPU with a batch size of 1. The base learning rate is set to 1×10^{-6}. The number of time steps for the diffusion model, T, is 1000. Experiments without and with patch-based fine-tuning are created after 800 and $10,000$ optimization steps, respectively. Unless otherwise noted, we adopt other hyperparameter choices from Stable Diffusion [39], and the results are generated with image resolution 512×512 and with latent dimension 64×64. For sampling parameters, we choose $K = 400$ and $v = 0.7$.

4.1. Qualitative Evaluation

To better understand various approaches, we collect images from a wide range of domains, *i.e.*, free-to-use high-resolution images from Flickr[1] and Unsplash[2]. During fine-

[1]https://www.flickr.com/
[2]https://unsplash.com/

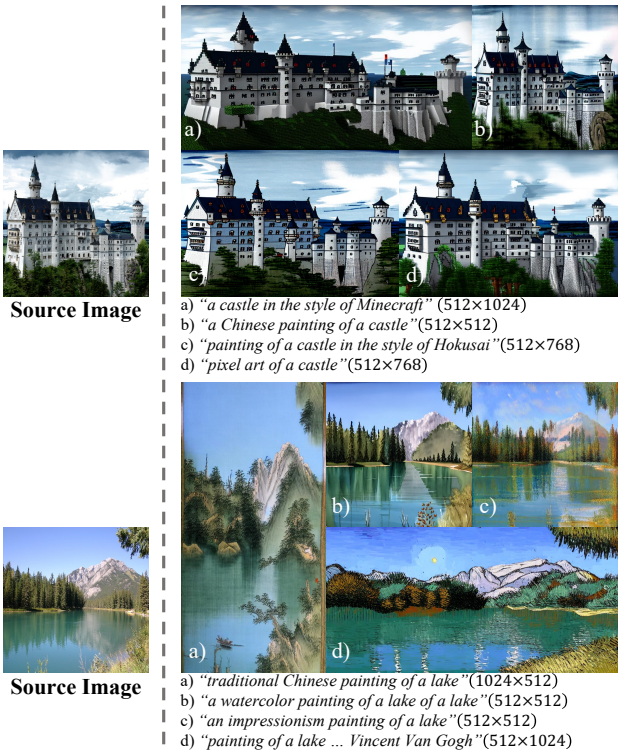


Figure 4. **Arbitrary resolution editing.** Our method achieves higher-resolution image editing without artifacts like duplicates, even on ones that change the height-width ratio drastically.

tuning, we apply a coarse class descriptor to the content we want to preserve, *e.g.*, dog, cat, castle, *etc*. After optimization, we edit each image with diverse editing prompts. We randomly generate 4 edit results for each image and editing prompt and choose the best one (such a process is also applied to other comparison methods). Our work shows impressive editing ability when applied to various images with different language guidance.

As presented in Fig. 3, using the model-based classifier-free guidance (Sec. 3.2) enables us to apply various editing via text prompts on the *single* real images. Each image has two text prompts describing different features we want to edit, *e.g.*, image style, background content, the texture of the content, *etc*. Our method can edit the related features while keeping the content intact. We further show our editing results on arbitrary resolution generation in Fig. 4. For each source image, we edit it with different prompts at various resolutions. As can be seen, our patch-based fine-tuning schedule (Sec. 3.3) successfully preserves the original portion and geometry features of the single source image, even on highly challenging resolution such as 512×1024.

4.2. Comparisons

We compare our method to concurrent leading techniques, Textual-Inversion [10] and DreamBooth [40], that can be used for single-image editing. Considering no official implementation has been released for DreamBooth, we adopt an unofficial but well-adopted and highly competitive implementation based on Stable Diffusion [39, 51]. We compare these techniques strictly according to the detailed guidance provided with the implementations.

Fig. 5 shows the comparison results. As can be noticed, our method maintains the fidelity of the images while applying changes as desired. Furthermore, our approach has high authenticity and structural integrity even for higher-resolution editing. For example, in the last row of Fig. 5, when the target prompt is "... *standing on grass*", our method generates results by modifying the texture of the land on which the dog stands with other features intact. However, other methods result in a dramatic change in the structure of the whole image. Moreover, in the second row, when modifying the painting *Mona Lisa*, both DreamBooth [40] and Textual-Inversion [10] fail to edit the image. Our work also shows clear advantages over the approaches on training-free editings, such as ILVR [7], SDEdit [32], and Prompt-to-prompt [20], with the qualitative comparisons presented in the *supplementary materials*.

4.3. Ablation Analysis

Patch-based fine-tuning. In Fig. 5, we show the results of editing images in higher resolution when fine-tuned without or with the proposed patch-based fine-tuning technique (w/o pos *vs*. w/ pos). When sampling at a higher resolution, as in the right part of Fig. 5, the denoising model fine-tuned without the patch-based training mechanism performs poorly. In the first row, the castle towers get duplicated to meet the resolution, and in the third row, the bench gets stretched disproportionately. In essence, our patch-based fine-tuning technique enables the diffusion model to leverage the super-resolution ability of the decoder and edit images at arbitrary resolution during testing time.

Analysis of model-based classifier-free guidance. We generate editing results by directly sampling from the fine-tuned model using Eqn. 2. We fine-tune the model without the patch-based schedule for 800 steps to retain more generalization ability. In this case, the model can perfectly reconstruct the source image while preserving as much editing ability as possible. We denote the setting as *w/o gudiance*. Using the same fine-tuned model, we conduct experiments under model-based classifier-free guidance, denoted as *w/ guidance*. As shown in Fig. 6, sampling without our model-based classifier-free guidance fails to react to the prompt, while our method can successfully edit images to match the target language guidance. We further analyze K and v in model-based classifier-free guidance.

Analysis on guidance step K in Sec. 3.2. In Fig. 7, we show our results on editing with different settings of K. We conduct this set of experiments by editing one single

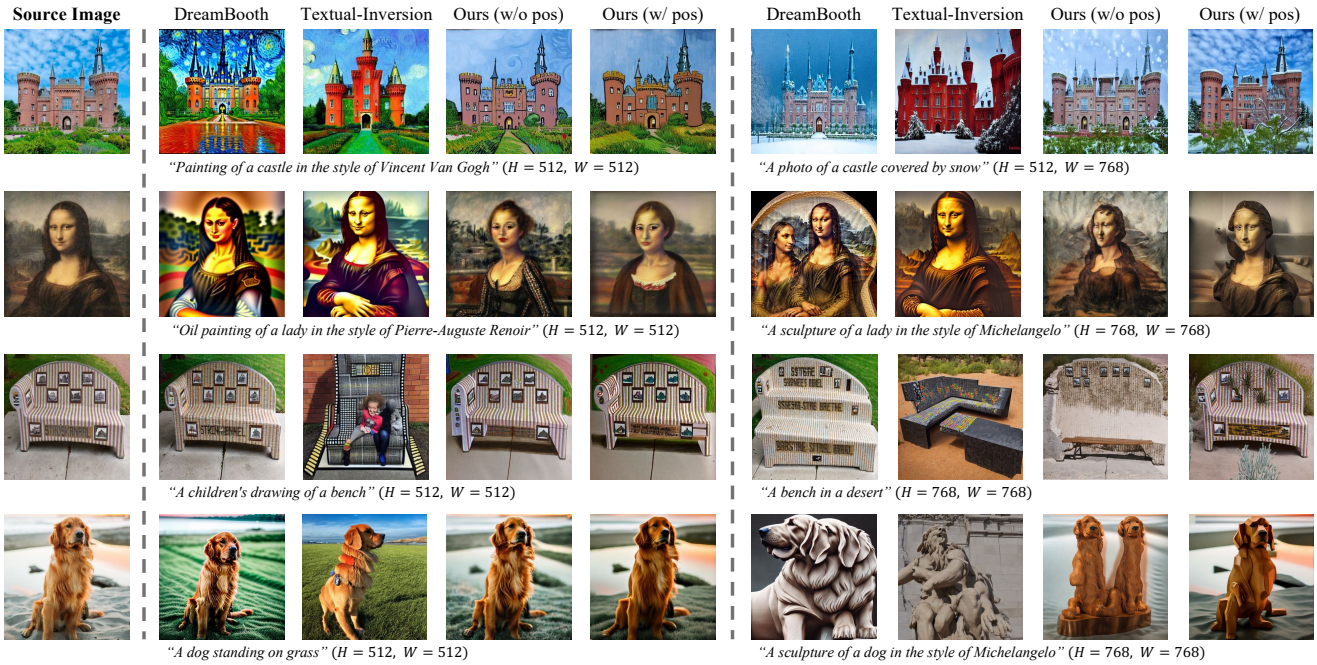


Figure 5. **Comparisons of various methods.** We compare our method to DreamBooth [40] and Textual-Inversion [10]. We adopt default hyper-parameters for both methods provided in their paper or code. On the left part of the figure, we edit at the resolution *same* as training time. On the right part, we edit the source image at a *higher* resolution. Our work successfully edits the image as required while preserving the details of the source images. We also compare our method without and with the patch-based fine-tuning mechanism (w/o pos *vs.* w/ pos). When editing at a fixed resolution, two settings perform equally, while at a higher resolution, the patch-based fine-tuning method successfully prevents artifacts.

image with the same language guidance at 768×768 resolution. We set $v = 0.7$. When $K = 0$, the model-based classifier-free guidance is applied for each step of the denoising process. Since the generalization ability of the fine-tuned model is limited, in this case, the model fails to apply the desired property to the single source image. When $K = 1,000$, the model-based classifier-free guidance is not applied to any step. Thus, the structure of the image is not preserved, and the generated result becomes a random sample of the pre-trained model.

We further show the quantitative results Fig. 8a. We repeat the abovementioned procedure over different K and randomly sample 20 images for each K. We calculate two metrics. To understand the editing result, the *image fidelity* that is measured by the LPIPS [57] distance between the original and the edited image. The *text alignment* calculated by the CLIP [36] score to understand the alignment between our generated images and target text. As can be seen, the image fidelity drops with the increase of K, indicating more details provided by the pre-trained model instead of the fine-tuned one. The text alignment measurement improves since the more details generated by the pre-trained model, the better editing results align with the target domain. To preserve the edit result's authenticity and fidelity to the source image, we set K as 400.

Analysis on guidance weight v in Sec. 3.2. We further study the impact of the guidance weight (v) in Fig. 9. We set $K = 400$ and resolution as 768×768 for each edit and use the same random seed to generate the result. As can be observed, the value of v controls the fidelity of the edit result. However, since the pre-trained model is trained at the resolution 512×512, the generated image contains many artifacts. When $v = 1$, the synthesized image entirely depends on the results from the pre-trained model. Additionally, we conduct quantitative experiments with LPIPS score measuring the image fidelity and CLIP score for the text alignment in Fig. 8b. When v is close to 1, the fidelity decreases while the edited feature decreases. When v is close to 0, the model relies mainly on the fine-tuned model for the output when $t > K$. However, since the fine-tuned model contains poor generalization ability, there is a significant amount of artifacts in the generated results, which leads to a poor LPIPS score. We choose 0.7 for each edit in this work as a trade-off between fidelity and creativity.

Other analysis. We defer the analysis of model-based classifier-free guidance with multiple inputs and fine-tuning without coarse class descriptor or prior-preservation loss to the *supplementary materials*.

Figure 6. **Analysis of model-based classifier-free guidance.** Directly sampling with target text using the fine-tuned model (w/o guidance) fails to generate images corresponding to the text prompt. In contrast, the model-based classifier-free guidance (w/ guidance) can synthesize high-fidelity images.

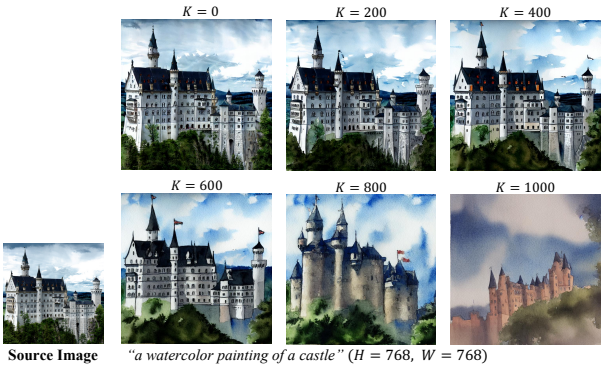


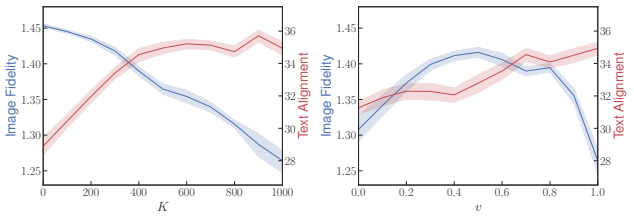
Figure 7. **Analysis on guidance step K.** Varying K, we can decide the steps where the model-based guidance is applied, which controls the details from the source image and edits to be applied.

5. Conclusion

This work introduces SINE, a method for single-image editing. With only one image and a brief description of the object in the image, our approach can enable a wide range of editing for arbitrary resolution, followed by the information depicted in the language guidance. To achieve such results, we leverage the pre-trained large-scale text-to-image diffusion model. Specifically, we first fine-tune the pre-trained model with our patch-based fine-tuning method until it overfits the single image. Then, during sampling time, we use the overfitted model to guide the pre-trained diffusion model for image synthesis, which maintains the fidelity of the results while taking advantage of the generalization ability of the pre-trained model. Compared with other methods, our approach has a better geometrical understanding of the image and thus can conduct complex editing to the images besides style transfer.

However, in some cases where confusing editing guidance is given for the diffusion model, *e.g.*, a chair-shaped dog, our method could fail. In cases where drastic changes are to be applied, *e.g.*, changing a dog to a tiger in the same posture, there are also noticeable artifacts. We show more examples in the *supplementary materials*.

One future direction is improving the fidelity of the editing results, which could be achieved by alleviating the overfitting problem of the fine-tuned model.



(a) Number of steps of the guidance. (b) Weight of the guidance.

Figure 8. **Trade-off between fidelity and alignment with target text.** We calculate the CLIP score [36] (a higher CLIP score indicates a better alignment between the edit result and target text) and 1 − LPIPS score [57] (showing the fidelity between edit results and source image, the higher, the better).

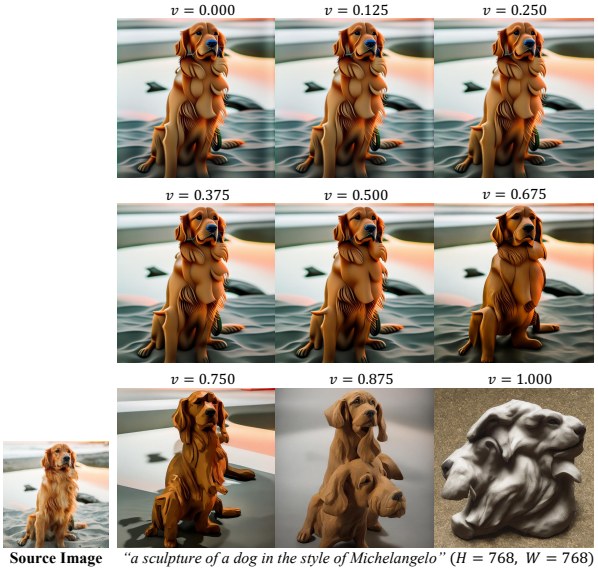


Figure 9. **Analysis of score interpolation.** Varying v with the same random seed gets editing results with various qualities.

References

[1] 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. 2

[2] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021. 2

[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

[4] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *ECCV*, pages 707–723. Springer, 2022. 1

[5] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. *ECCV*, 2022. 2, 4

[6] Yuxiao Chen, Jianbo Yuan, Yu Tian, Shijie Geng, Xinyu Li, Ding Zhou, Dimitris N. Metaxas, and Hongxia Yang. Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

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

[8] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 3

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 4

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 4, 6, 7

[11] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *TOG*, 41(4):1–13, 2022. 1, 2

[12] Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. Hiclip: Contrastive language-image pretraining with hierarchy-aware attention. *arXiv preprint arXiv:2303.02995*, 2023. 2

[13] Niv Granot, Ben Feinstein, Assaf Shocher, Shai Bagon, and Michal Irani. Drop the gan: In defense of patches nearest neighbors as single image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13460–13469, 2022. 4

[14] Ligong Han, Ruijiang Gao, Mun Kim, Xin Tao, Bo Liu, and Dimitris Metaxas. Robust conditional gan from uncertainty-aware pairwise comparisons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10909–10916, 2020. 2

[15] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 1

[16] Ligong Han, Martin Renqiang Min, Anastasis Stathopoulos, Yu Tian, Ruijiang Gao, Asim Kadav, and Dimitris N Metaxas. Dual projection generative adversarial networks for conditional image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14438–14447, 2021. 2

[17] Ligong Han, Sri Harsha Musunuri, Martin Renqiang Min, Ruijiang Gao, Yu Tian, and Dimitris Metaxas. Ae-stylegan: Improved training of style-based auto-encoders. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3134–3143, 2022. 2

[18] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3615–3625, 2022. 2

[19] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020. 2

[20] 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. 1, 2, 4, 6

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NIPS*, 33:6840–6851, 2020. 2, 3

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2, 3

[23] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 4

[24] 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. 2

[25] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, pages 2426–2435, 2022. 1, 2

[26] Hyunsu Kim, Yunjey Choi, Junho Kim, Sungjoo Yoo, and Youngjung Uh. Exploiting spatial dimensions of latent in gan for real-time image editing. In *CVPR*, pages 852–861, 2021. 1

[27] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 2

[28] Jason Lee, Kyunghyun Cho, and Douwe Kiela. Countering language drift via visual grounding. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 4385–4395. Association for Computational Linguistics, 2020. 2

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

[30] Zhi-Song Liu, Li-Wen Wang, Wan-Chi Siu, and Vicky Kalogeiton. Name your style: An arbitrary artist-aware image style transfer. *arXiv preprint arXiv:2202.13562*, 2022. 2

[31] Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. Countering language drift with seeded iterated learning. In *ICML*, pages 6437–6447. PMLR, 2020. 2

[32] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 1, 2, 6

[33] 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. 2

[34] Byong Mok Oh, Max Chen, Julie Dorsey, and Frédo Durand. Image-based modeling and photo editing. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 433–442, 2001. 2

[35] 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. 2

[36] 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. 2, 7, 8

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

[38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 3, 4, 5, 6

[40] 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. 1, 2, 3, 4, 6, 7

[41] Dohoon Ryu and Jong Chul Ye. Pyramidal denoising diffusion probabilistic models. *arXiv preprint arXiv:2208.01864*, 2022. 2

[42] 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. *NIPS*, 2022. 1, 2, 3

[43] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5

[44] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4570–4580, 2019. 4

[45] Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. Ingan: Capturing and retargeting the" dna" of a natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4492–4501, 2019. 4

[46] 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. 2

[47] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2

[48] Kunpeng Song, Ligong Han, Bingchen Liu, Dimitris Metaxas, and Ahmed Elgammal. Diffusion guided domain adaptation of image generators. *arXiv preprint arXiv:2212.04473*, 2022. 2

[49] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[50] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020. 2

[51] Xavierxiao. Xavierxiao/dreambooth-stable-diffusion: Implementation of dreambooth (https://arxiv.org/abs/2208.12242) with stable diffusion. 6

[52] Rui Xu, Xintao Wang, Kai Chen, Bolei Zhou, and Chen Change Loy. Positional encoding as spatial inductive bias in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13569–13578, 2021. 2

[53] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference*

on computer vision and pattern recognition, pages 1316–1324, 2018. 2

[54] Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *arXiv preprint arXiv:2107.02423*, 2021. 2

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

[56] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021. 2

[57] 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, 8

[58] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *ECCV*, pages 592–608. Springer, 2020. 1

[59] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 1

[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. 2