Uncovering the Disentanglement Capability in Text-to-Image Diffusion Models

Qiucheng Wu[1], Yujian Liu[1], Handong Zhao[2], Ajinkya Kale[2], Trung Bui[2], Tong Yu[2],
Zhe Lin[2], Yang Zhang[3], Shiyu Chang[1]
[1]UC, Santa Barbara, [2]Adobe Research, [3] MIT-IBM Watson AI Lab
{qiucheng, yujianliu}@ucsb.edu

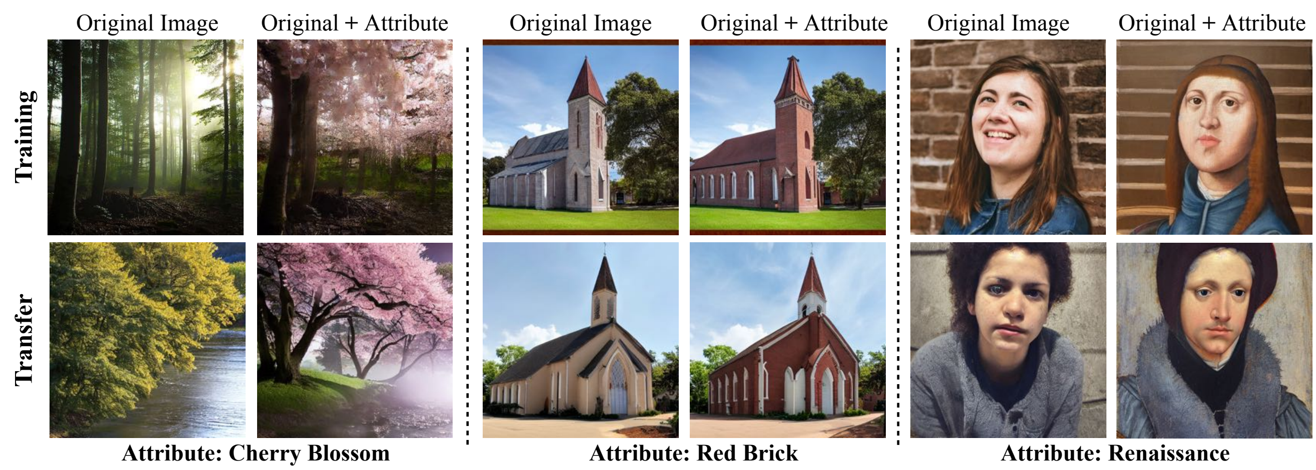

Figure 1. Example attributes disentangled from the stable diffusion model. Based on a *fixed* stable diffusion model, we disentangle the target attribute from a single training image. The learned parameters can then be applied to an unseen image and achieve the same edit.

Abstract

*Generative models have been widely studied in computer vision. Recently, diffusion models have drawn substantial attention due to the high quality of their generated images. A key desired property of image generative models is the ability to disentangle different attributes, which should enable modification towards a style without changing the semantic content, and the modification parameters should generalize to different images. Previous studies have found that generative adversarial networks (GANs) are inherently endowed with such disentanglement capability, so they can perform disentangled image editing without re-training or fine-tuning the network. In this work, we explore whether diffusion models are also inherently equipped with such a capability. Our finding is that for stable diffusion models, by partially changing the input text embedding from a neutral description (*e.g.*, "a photo of person") to one with style (*e.g.*, "a photo of person with smile") while fixing all the Gaussian random noises introduced during the denoising process, the generated images can be modified towards the target style without changing the semantic content. Based on this finding, we further propose a simple, light-weight image editing algorithm where the mixing weights of the two text embeddings are optimized for style matching and content preservation. This entire process only involves optimizing over around 50 parameters and does not fine-tune the diffusion model itself. Experiments show that the proposed method can modify a wide range of attributes, with the performance outperforming diffusion-model-based image-editing algorithms that require fine-tuning. The optimized weights generalize well to different images. Our code is publicly available at* https://github.com/UCSB-NLP-Chang/DiffusionDisentanglement.

1. Introduction

Image generation has been a widely-studied research problem in computer vision, with many competitive generative models proposed over the last decade, such as generative adversarial networks (GANs) [5, 10, 18, 30, 32] and variational autoencoders (VAE) [39, 57, 59, 60]. Recently, diffusion models [23, 71–73], with their ability to generate high-quality and high-resolution images in different do-

		Scenes	Person
✔	Global	**Styles** (children drawing, cyberpunk, anime), **Building appearance** (wooden, red brick), **Weather & time** (sunset, night, snowy)	**Styles** (renaissance, Egyptian mural, sketch, Pixar) **Appearance** (young, tanned, male)
	Local	Cherry blossom, rainbow, foothills	**Expressions** (smiling, crying, angry)
✘	Small edits	Cake toppings, remove people on the street	Hats, hair colors, earrings

Table 1. Summarization of explored attributes. ✔ shows successfully disentangled attributes and ✘ shows failure cases. Small edits on the image are harder to be disentangled when the target attribute correlates with other parts of the image.

mains, have soon attracted wide research attention.

One important research direction regarding image generative models is the ability to *disentangle* different aspects of the generated images, such as semantic contents and styles, which is crucial for image editing and style transfer. A generative model with a good disentanglement capability should satisfy the following two desirable properties. First, it should permit separate modification of one aspect *without* changing other aspects. As an example shown in Fig. 2, in text-to-image generation, when the text input changes from "a photo of person" to "a photo of person with smile", the generative model should have the ability to modify just the expression of the person (*i.e.,* from the top image to middle image in Fig. 2) without changing the person's identity (the bottom image in Fig. 2). Second, the parameters learned from modifying one image should transfer well to other similar images. For example, the optimal parameters that can add smile to one person should also work for images of different people with different genders and races.

Previous studies have discovered that GANs are inherently endowed with a strong disentanglement capability. Specifically, it is found that there exist certain directions in the latent space separately controlling different attributes. Therefore, by identifying these directions, *e.g.,* via principal component analysis [19], GAN can achieve effective disentanglement *without* any re-training or fine-tuning. On the other hand, such an inherent disentanglement capability has yet to be found in diffusion models. Hence come our research questions: Do diffusion models also possess a disentanglement capability with the aforementioned nice properties? If so, how can we uncover it?

In this paper, we seek to answer these research questions. Our finding is that for stable diffusion model [61], one of the diffusion models that can generate images based on an input text description, disentangled image modifications can be achieved by partial modifications in the text embedding space. In particular, if we fix the standard Gaussian noises introduced in the denoising process, and partially change the input text embedding from a neutral description (*e.g.,* "a photo of person") to one with style (*e.g.,* "a photo of person with smile"), the generated image will also shift towards the target style without changing the semantic content. Based on this finding, we further propose a simple, light-weight algorithm, where we optimize the mixing weights of the two text embeddings under two objectives, a perceptual loss for content preservation and a CLIP-based style matching loss. The entire process only involves optimizing over around 50 parameters and does not fine-tune the diffusion model.

Our experiments show that the inherent disentanglement capability in stable diffusion model can already disentangle a wide range of concepts and attributes, ranging from global styles such as painting styles to local styles like facial expressions, as shown in Table 1. As shown in Fig. 1, by learning the optimal mixing weights of the two descriptions, stable diffusion models can generate convincing image pairs that only modify the target attribute, and the optimal weights can generalize well to different images. The experiment results also show that our proposed image editing algorithm, without fine-tuning the diffusion model, can match or outperform the more sophisticated diffusion-model-based image-editing baselines that require fine-tuning. The findings of this paper can shed some light on how diffusion models work and how they can be applied to image editing tasks.

2. Related Works

Disentanglement in Generative Models: The ability to disentangle different attributes is a key desired property of generative models. Previous work that studies disentanglement mainly aims to learn parameters that allow modifications on a target aspect without changing other aspects, and the learned parameters should generalize to different images [3, 15, 32]. For pre-trained GANs [10, 31–33], it has been shown that the disentanglement can be achieved by moving towards particular directions in its latent space [19, 68–70], which will lead to attribute-only changes [1, 2, 53]. Multiple methods have been proposed to discover these latent directions, which leverage auxiliary classifiers [68, 69], principal component analysis [19] , contrastive learning [58], and information maximization [12]. Besides GANs, disentanglement has also been studied in VAE and flow-based models [36, 52]. Recently, two works study disentanglement in diffusion models. The first work disentangles attributes by learning a shift in the embedding space of an intermediate layer of U-Net [41, 62], such that applying the shift satisfies the disentanglement criteria. Particularly, they use a neural network to generate such shifts. However, disentangling in the hidden layer representation of U-Net might be sub-optimal, as can be observed that their method struggles at disentangling holistic styles of the image. Moreover, their search space and number of parameters are much larger than ours. By contrast, we consider the text embedding space, which is more natural for text-to-image diffusion models and achieves comparable or better results with only 1.2% parameters of theirs. Another work that studies disentanglement is [54]. They train an encoder to generate an image-specific representation, which is later used as input to diffusion models to reconstruct the original image. Disentanglement is done by finding corresponding directions in this representation space similar to methods in GANs [68, 69]. However, their method requires retraining a diffusion model from scratch, whereas we fix the pre-trained diffusion model.

Diffusion Models: Diffusion models [23,71–73] are a family of generative models that have achieved state-of-the-art performance in image synthesis and have advanced research in super-resolution [24, 66], inpainting [49, 64], density estimation [38], video synthesis [22, 26], and areas beyond computer vision [6, 11, 27, 40, 74]. Building on top of diffusion models, various methods have been proposed to control the generation process through external models [14, 46] or additional inputs [25]. One type of conditional generation models is the text-to-image diffusion models [51,56,61,65], which take text descriptions as inputs and generate images that match the text descriptions. Due to the expressiveness of text and superior generation quality of diffusion models, these models allow unprecedented control over generated images and have inspired many novel applications.

Image Editing: Image editing is a widely-studied task [21, 77]. Many GAN-based editing works [9, 28, 42, 47, 53] have demonstrated strong controllability. Recently, diffusion models have been broadly adapted to image editing task [13, 20, 43, 45, 50]. With the CLIP encoder [55] that bridges text and image, generation process can be guided by arbitrary text descriptions [46]. To preserve the contents in a local region, [8] relies on an auxiliary mask, such that contents in the unmasked region are largely kept unchanged during generation. Moreover, some works [16, 34, 63] propose to invert the input image to find text embeddings that can synthesize the same object but in different scenes and views. Although these works have demonstrated successful edits, there are two limitations. First, most of the methods require fine-tuning diffusion models [34, 35, 63]. For each editing task, they have to fine-tune and store the whole diffusion model, making them unscalable to a large amount of edits. Second, many methods rely on auxiliary inputs such as image masks [4, 7, 8] or multiple examples of the edited object [16, 63], which are not always available. Besides, only editing masked region may cause incoherence between masked and unmasked regions. In this work, leveraging the disentanglement in stable diffusion, we propose to perform image editing without auxiliary inputs and the need to fine-tune diffusion models, which is more practical to use.

3. Attribute Disentanglement in Stable Diffusion Models

In this section, we will explore the disentanglement properties inherent in diffusion models, and then propose an approach for disentangled image modification and editing utilizing these properties.

3.1. Preliminaries on Diffusion Models

We first provide a brief overview of the denoising diffusion implicit model (DDIM) [72] conditioned on input text descriptions that is used in the stable diffusion model [61], which we will be primarily studying in this paper. Given a text embedding, denoted as $\boldsymbol{c}$, the goal of the text-conditioned DDIM is to generate an image, denoted as $\boldsymbol{X}_0$, that conforms to the text description. DDIM defines a set of noisy images, $\boldsymbol{X}_{1:T}$, by adding Gaussian noise to $\boldsymbol{X}_0$ according to a predefined diffusion process. Each $\boldsymbol{X}_t$ is corrupted with a larger noise than $\boldsymbol{X}_{t-1}$, and $\boldsymbol{X}_T$ is very close to standard Gaussian noise. The generation process of DDIM tries to denoise from $\boldsymbol{X}_T$ all the way back to $\boldsymbol{X}_0$. Specifically, in the first step, $\boldsymbol{X}_T$ is randomly drawn from a standard Gaussian distribution. Then, each $\boldsymbol{X}_{t-1}$ is inferred from $\boldsymbol{X}_t$ via the following denoising process:

$$\boldsymbol{X}_{t-1} = \gamma_{t0}\boldsymbol{X}_t + \gamma_{t1}\boldsymbol{\epsilon}_\theta(\boldsymbol{X}_t, t, \boldsymbol{c}_t), \tag{1}$$

where $\boldsymbol{c}_t$ is the text embedding used at step t. In most common settings, generating one image only requires one text description $\boldsymbol{c}$, so $\boldsymbol{c}_t = \boldsymbol{c}$ for all t. Here we make it dependent upon t to accommodate the discussions in the following subsections. $\boldsymbol{\epsilon}_{\theta^*}(\boldsymbol{X}_t, t, \boldsymbol{c}_t)$ is a pre-trained denoising network that infers $\boldsymbol{X}_0$ given the input of $\boldsymbol{X}_t$ and $\boldsymbol{c}_t$. The parameters of the denoising network, θ^*, are considered fixed throughout this section. γ_{t0} and γ_{t1} are defined as

$$\gamma_{t0} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}, \quad \gamma_{t1} = \sqrt{1-\alpha_{t-1}} - \sqrt{\frac{\alpha_{t-1}}{\alpha_t} - \alpha_{t-1}}, \tag{2}$$

and $\alpha_{0:T}$ are hyperparameters that govern diffusion process.

Note that in the general DDIM framework, a Gaussian noise is added to each denoising step in Eq. (1), but we follow the convention [35] to set the variance to 0 for better stability. Therefore, the generated image is a *deterministic* function of initial random noise $\boldsymbol{X}_T$ and the text descriptions $\boldsymbol{c}_{1:T}$. Thus we introduce the following notation,

$$\boldsymbol{X}_0 = \boldsymbol{g}(\boldsymbol{X}_T, \boldsymbol{c}_{1:T}), \tag{3}$$

to summarize the image generation process. The stable diffusion model follows the same setup, except that the diffusion process is performed on a hidden embedding space, so Eq. (3) can also summarize its generation process.

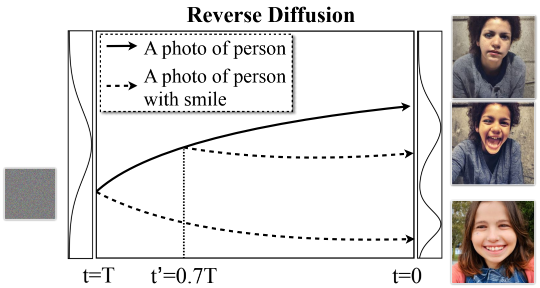

Figure 2. The disentanglement property of stable diffusion models. The top image is generated conditioned on "a photo of person". The bottom image is generated with all descriptions replaced with "a photo of person with smile", and changes the person's identity. The middle image is generated by partially replacing descriptions at later steps, and maintains the person's identity.

3.2. The Disentanglement Properties

Now we are ready to study whether the stable diffusion model is inherently capable of disentangling styles from semantic content. For concreteness, we will present our findings based on one specific example, but the findings are consistent across different cases.

Consider two text embeddings, $\boldsymbol{c}^{(0)}$ and $\boldsymbol{c}^{(1)}$. $\boldsymbol{c}^{(0)}$ is the embedding of a style-neutral description, "a photo of person", and $\boldsymbol{c}^{(1)}$ is the embedding of a description with an explicit style, "a photo of person with smile". When $\boldsymbol{c}^{(0)}$ is fed to the model, the generated image is a person with neutral expressions (top image in Fig. 2). We investigate whether the model can generate an image of the *same person* with only the facial expression changed, when we fix $\boldsymbol{X}_T$ (hence controlling all the randomness) and replace $\boldsymbol{c}^{(0)}$ with $\boldsymbol{c}^{(1)}$.

Case 1: Full Replacement. In our first attempt, we replaced text embeddings at all denoising steps with $\boldsymbol{c}^{(1)}$. The resulting generated image undesirably changes the identity of the person, as shown in the bottom image in Fig. 2.

Case 2: Partial Replacement. In our second attempt, we replaced the text embeddings to $\boldsymbol{c}^{(1)}$ only at the later denoising steps, *i.e.,*

$$\boldsymbol{c}_t = \boldsymbol{c}^{(1)}, \forall t \leq t', \quad \boldsymbol{c}_t = \boldsymbol{c}^{(0)}, \forall t > t', \tag{4}$$

where $t' = 0.7T$ in this example. In this case, the image can successfully maintain the identity of the person while only changing the smile, as shown in the middle image in Fig. 2.

As can be concluded, there exists an inherent disentanglement capability in stable diffusion models, which can be triggered by partially replacing the text embeddings, but not with full replacement. We have performed the same experiments on more objects and styles and the observations are consistent, as shown in Appendix C.

3.3. Optimizing for Disentanglement

In Sec. 3.2, we have seen one way of realizing a disentangled generation of image using $\boldsymbol{c}^{(0)}$ and $\boldsymbol{c}^{(1)}$. However, this is not necessarily the optimal way of combining the two embeddings in terms of disentanglement. In this subsection, we will propose a principled and tractable optimization scheme to combine a given pair of $\boldsymbol{c}^{(0)}$ and $\boldsymbol{c}^{(1)}$ to achieve the best disentanglement, illustrated in Fig. 3.

The key relaxation for the optimization framework is the *soft combination* of the text embeddings. Specifically, instead of feeding either $\boldsymbol{c}^{(0)}$ or $\boldsymbol{c}^{(1)}$ at each denoising step t, we feed a soft combination of the two, namely

$$\boldsymbol{c}_t = \lambda_t \boldsymbol{c}^{(1)} + (1 - \lambda_t)\boldsymbol{c}^{(0)} \equiv \boldsymbol{c}_t^{(\lambda)}, \tag{5}$$

where λ_t is a learnable combination weight. The soft combination offers a much richer representation power, with both cases discussed in Sec. 3.2 being its special cases (by setting λ_t to either 1 or 0).

Given a random noise $\boldsymbol{X}_T$ and the text description pair $\boldsymbol{c}^{(0)}$ and $\boldsymbol{c}^{(1)}$, the optimization procedure for $\lambda_{1:T}$ is as follows. First, two images are generated, a style-neutral one generated with $\boldsymbol{c}^{(0)}$ and the other generated with $\boldsymbol{c}_{1:T}^{(\lambda)}$:

$$\boldsymbol{X}_0^{(0)} = \boldsymbol{g}(\boldsymbol{X}_T, \boldsymbol{c}^{(0)}), \quad \boldsymbol{X}_0^{(\lambda)} = \boldsymbol{g}(\boldsymbol{X}_T, \boldsymbol{c}_{1:T}^{(\lambda)}). \tag{6}$$

Our goal is to find an optimal $\lambda_{1:T}$ such that $\boldsymbol{X}_0^{(\lambda)}$ maintains the same semantic content as $\boldsymbol{X}_0^{(0)}$ but conforms to the style described in $\boldsymbol{c}^{(1)}$, which is achieved by solving the following optimization problem similar to [41, 53]:

$$\min_{\lambda_{1:T}} \mathcal{L}_{\text{clip}}(\boldsymbol{X}_0^{(0)}, \boldsymbol{X}_0^{(\lambda)}, \boldsymbol{c}^{(0)}, \boldsymbol{c}^{(1)}) + \beta \mathcal{L}_{perc}(\boldsymbol{X}_0^{(0)}, \boldsymbol{X}_0^{(\lambda)}). \tag{7}$$

β is the hyperparameter that balances the two loss terms. $\mathcal{L}_{\text{clip}}$ is the directional CLIP loss [17] that encourages the image change from $\boldsymbol{X}_0^{(0)}$ to $\boldsymbol{X}_0^{(\lambda)}$ matches the text change from $\boldsymbol{c}^{(0)}$ to $\boldsymbol{c}^{(1)}$ in the CLIP embedding space.[1] $\mathcal{L}_{perc}$ is the perceptual loss [29] that prevents drastic changes in semantic content:

$$\mathcal{L}_{perc}(\boldsymbol{X}_0^{(0)}, \boldsymbol{X}_0^{(\lambda)}) = \|\boldsymbol{h}(\boldsymbol{X}_0^{(0)}) - \boldsymbol{h}(\boldsymbol{X}_0^{(\lambda)})\|_1, \tag{8}$$

where $\boldsymbol{h}(\cdot)$ denotes a perceptual network that encodes a given image. In stable diffusion models, the number of denoising steps can be as few as 50 and so are the corresponding $\lambda_{1:T}$, so this optimization problem involves very few parameters and does not need to fine-tune the diffusion model itself.

3.4. Extension to Image Editing

With the disentangled image modification algorithm developed, we can now extend the approach to achieve disentangled image editing. The only difference of the image editing setting compared to the settings in the previous subsections is that rather than having the diffusion model generate the neutral image conditioned on $\boldsymbol{c}^{(0)}$, the neutral image

[1] In practice, we use a different text embedder for $\mathcal{L}_{clip}$, but for brevity we still use $\boldsymbol{c}^{(0)}$ and $\boldsymbol{c}^{(1)}$ to denote the embeddings.

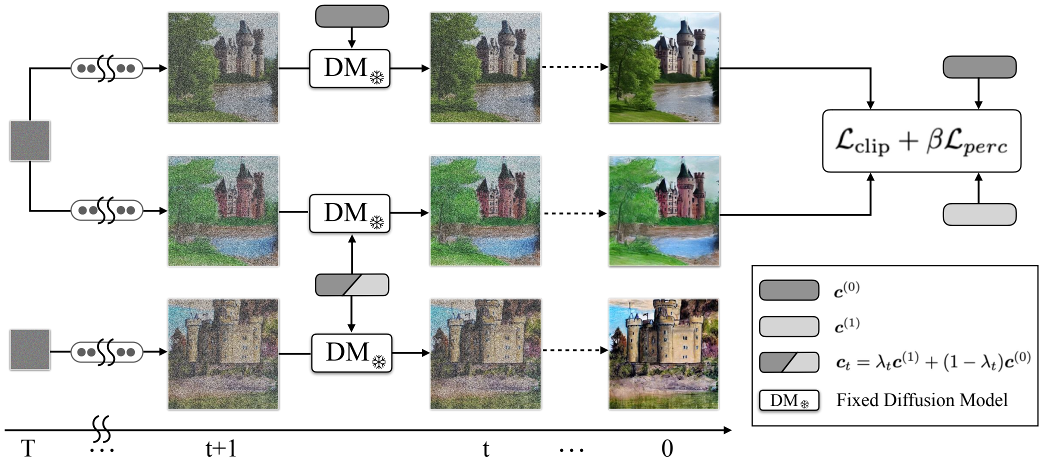

Figure 3. Overview of our method that finds optimal text embedding for disentanglement. In this example, $\boldsymbol{c}^{(0)}$ is the embedding of "A castle", and $\boldsymbol{c}^{(1)}$ is the embedding of "A children drawing of castle". **First two rows:** optimization process that finds the best soft combination of $\boldsymbol{c}^{(0)}$ and $\boldsymbol{c}^{(1)}$, such that the modified image (the second row) changes the attribute without affecting other contents. **Last row:** the learned text embedding can be directly applied to a new image, which leads to the same editing effect.

is now externally given, which we denote as $\boldsymbol{I}$ to show the distinction. Therefore, if we could find a value for initial random variable $\boldsymbol{X}_T$ such that the stable diffusion model can generate exactly the same image as $\boldsymbol{I}$ when conditioned on $\boldsymbol{c}^{(0)}$, we can then use the same approach in Sec. 3.3 for the image editing task.

To this end, we adapt the image inversion approach proposed in [20, 35], which recursively generates a set of noisy images, $\hat{\boldsymbol{X}}_{1:T}$ based on $\boldsymbol{I}$ as follows:

$$\hat{\boldsymbol{X}}_0 = \boldsymbol{I}, \quad \hat{\boldsymbol{X}}_{t+1} = \gamma'_{t0}\hat{\boldsymbol{X}}_t + \gamma'_{t1}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\hat{\boldsymbol{X}}_t, t, \boldsymbol{c}^{(0)}), \tag{9}$$

where

$$\gamma'_{t0} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}, \quad \gamma'_{t1} = \sqrt{1-\alpha_{t+1}} - \sqrt{\frac{\alpha_{t+1}}{\alpha_t} - \alpha_{t+1}}. \tag{10}$$

It has been shown [35] that by setting $\boldsymbol{X}_T = \hat{\boldsymbol{X}}_T$ and following through the generation process in Eq. (1) with $\boldsymbol{c}_t = \boldsymbol{c}^{(0)}$, the resulting $\boldsymbol{X}_{0:T-1}$ would satisfy $\boldsymbol{X}_t \approx \hat{\boldsymbol{X}}_t, \forall t \in [0, T-1]$, and therefore $\boldsymbol{X}_0 \approx \boldsymbol{I}$.

To further close the approximation gap, we introduce a new diffusion process, where the approximation error is added as a correction term. Formally, the new diffusion process starts with $\tilde{\boldsymbol{X}}_T = \hat{\boldsymbol{X}}_T$, and then

$$\tilde{\boldsymbol{X}}_{t-1} = \gamma_{t0}\tilde{\boldsymbol{X}}_t + \gamma_{t1}\boldsymbol{\epsilon}_\theta(\tilde{\boldsymbol{X}}_t, t, \boldsymbol{c}_t) + \boldsymbol{E}_t, \tag{11}$$

where the $\boldsymbol{E}_t$ is the correction term defined as

$$\boldsymbol{E}_t = \hat{\boldsymbol{X}}_{t-1} - \gamma_{t0}\hat{\boldsymbol{X}}_t - \gamma_{t1}\boldsymbol{\epsilon}_\theta(\hat{\boldsymbol{X}}_t, t, \boldsymbol{c}^{(0)}). \tag{12}$$

Again, for notational brevity, we summarize this new generation process as

$$\tilde{\boldsymbol{X}}_0 = \tilde{\boldsymbol{g}}(\tilde{\boldsymbol{X}}_T, \boldsymbol{c}_{1:T}, \boldsymbol{E}_{1:T}). \tag{13}$$

It can be easily shown that $\tilde{\boldsymbol{g}}(\hat{\boldsymbol{X}}_T, \boldsymbol{c}^{(0)}, \boldsymbol{E}_{1:T}) = \boldsymbol{I}$.

Now that we have developed a generation process that can reconstruct $\boldsymbol{I}$, we can now follow the same procedure in Sec. 3.3 to perform image editing, with the image generation in Eq. (6) replaced with

$$\tilde{\boldsymbol{X}}_0^{(0)} = \tilde{\boldsymbol{g}}(\hat{\boldsymbol{X}}_T, \boldsymbol{c}^{(0)}, \boldsymbol{E}_{1:T}) = \boldsymbol{I}, \ \tilde{\boldsymbol{X}}_0^{(\lambda)} = \tilde{\boldsymbol{g}}(\hat{\boldsymbol{X}}_T, \boldsymbol{c}_{1:T}^{(\lambda)}, \boldsymbol{E}_{1:T}). \tag{14}$$

It is worth emphasizing that when generating $\tilde{\boldsymbol{X}}_0^{(\lambda)}$, the error correction terms $\boldsymbol{E}_{1:T}$ are still fixed to the ones computed for reconstructing $\boldsymbol{I}$. To further enhance the quality of the edited image, we adopt the re-diffusion approach in [50].

4. Experiments

We will perform experiments to explore the inherent disentanglement capability in the stable diffusion model and evaluate the performance of our method.

Implementation Details: For all experiments, we use the diffusion model `stable-diffusion-v1-4` [61], which is pre-trained on `laion` dataset [67]. The pre-trained model is *frozen* throughout all experiments, and we keep the default hyperparameters of the model. All images generated by our method are in size 512×512. We use a variant of the DDIM sampler [44] to synthesize images with 50 total backward diffusion steps. To optimize $\lambda_{1:T}$, we use Adam [37] optimizer with learning rate 0.03. To balance two loss terms, we set β to 0.05 for all human face experiments and 0.03 for all experiments on scenes and buildings. Finally, when editing real images, the number of re-diffusion steps is 20. More information on the hyperparameters and optimization is in Appendix A.

4.1. Exploring the Disentanglement Capability

Since the foundation of our method hinges on the inherent disentanglement capability in the stable diffusion

Figure 4. Example disentanglement results. Text description with style ($\boldsymbol{c}^{(1)}$) is shown below each row, which consists of the style-neutral text description ($\boldsymbol{c}^{(0)}$) and target attribute description, separated by comma. **Top two panels:** successful cases. **Bottom panel:** failure cases. Within each attribute, **first row:** results on optimization images; **second row:** results of transferring to unseen images; **left column:** source images; **right column:** modified images. More examples can be found in Appendix D.

model, we would like to first explore the strength and the limit of this capability. In particular, the research question we would like to answer is what are the attributes and objects that the stable diffusion model can inherently disentangle well, and what cannot. To this end, we perform a comprehensive qualitative study where we first compile a list of objects and attributes that have been tested upon in existing work on image disentanglement [19,41,69], as shown in Table 1. For each object-attribute pair in the list, we craft a text description pair, the neutral description $\boldsymbol{c}^{(0)}$ and the description with style $\boldsymbol{c}^{(1)}$. In particular, $\boldsymbol{c}^{(0)}$ includes just the name of the object, *e.g.,* "A castle", and $\boldsymbol{c}^{(1)}$ is constructed by appending the description of the attribute to $\boldsymbol{c}^{(0)}$, separated by comma (*e.g.,* "A castle, children drawing style"). Next, we generate five style-neutral images for each $\boldsymbol{c}^{(0)}$ and use the optimization method in Sec. 3.3 to disentangle the attribute from these five images (more details in Appendix B).

Results: The results are shown in Table 1, and some generated image pairs with their text descriptions are shown in Fig. 4, where the top two panels show some successful cases. In each panel, the first row (marked "Training") shows the results where the optimal $\lambda_{1:T}$ are learned specific to that image pair; the second row (marked "Transfer") shows the results of applying the learned $\lambda_{1:T}$ to a

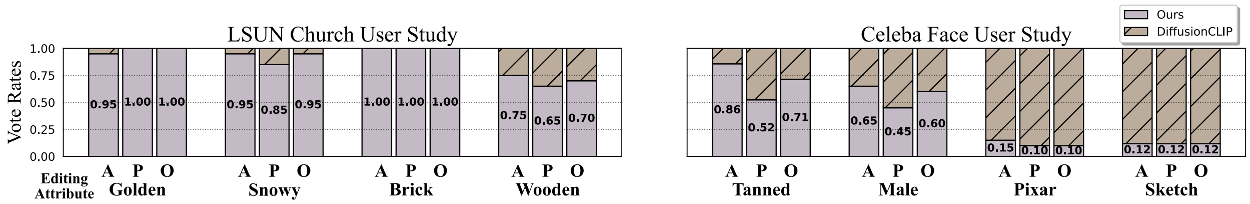

Figure 5. Human evlautions on **A (attribute similarity)**, **P (content preservation)**, and **O (overall quality)**. The number shown in each bar indicates the percentage of users who vote for our method is better than the baseline for the corresponding aspect.

new image with the same object and target attribute. As can be observed, using our method, the stable diffusion model, without any re-training, can disentangle a surprisingly wide range of objects and attributes already. In particular, our method is strong at disentangling global styles that cover largely the whole image, such as scenery styles, drawing styles, and architecture materials, where the scenery layout or building structure is largely maintained while only the target attribute is modified. Our method can also disentangle many local attributes like facial expressions. Besides, the learned $\lambda_{1:T}$ has a great transferability to unseen images, thus satisfying both disentanglement criteria in Sec. 1.

On the other hand, stable diffusion has difficulties disentangling attributes that involve small objects, *e.g.,* adding small accessories, as shown in the bottom panel of Fig. 4. When the target attribute is added, the model tends to also change some other correlated attributes, such as the style of the cake or the identity of the person, which may be ascribed to the model's weaker control of finer-grained details. Nevertheless, these results suggest that the disentanglement capability in an unmodified stable diffusion model may be stronger than previously revealed, which provides a good justification for our method that applies it as is to image editing. More examples are shown in Appendix D.

4.2. Evaluation on Disentangled Image Editing

In this section, we will evaluate our proposed method on the image editing task as described in Sec. 3.4.

Configurations: Since there is no ground-truth for the image editing task, we conduct a subjective evaluation on Amazon Mechanical Turk. Specifically, we use the Celeb-A [48] and LSUN-church [76]. For each dataset, we use the first 20 images as the source image and perform 4 types of the edit used in [35] (*i.e.,* tanned, male, sketch, pixar for human faces; and golden, wooden, snowy, red brick for churches). The main baseline is DiffusionCLIP, which is a state-of-the-art diffusion-based image editing approach that requires fine-tuning. Each subject was presented with a pair of images edited by the two methods from the same source image in a randomized order and asked three questions regarding the editing quality: (1) (*attribute similarity*) which image better incorporates the target attribute in a natural way; (2) (*content preservation*) which image better preserves other contents; and (3) (*overall*) which image has better overall quality. More details are in Appendix E.

Subjective Evaluation Results: Fig. 5 shows the results of the percentage of answers that chose our method in each of the three questions and in each target attribute. As can be observed, in 6 out of the 8 attributes, our method outperforms the baseline in almost all three aspects, demonstrating the high quality of its disentanglement. We discovered a common failure case in DiffusionCLIP where the attributes are so over-optimized that some artifacts are introduced and some irrelevant parts of the image are modified (examples in Appendix E). Our method, with its lightweight optimization, can avoid the over-optimization problem. On the other hand, our method is less competitive in human-related editing, underperforming the baseline in two attributes, "sketch" and "pixar". One potential cause is that the diffusion model in DiffusionCLIP is specifically trained to generate human images. Nevertheless, these results show that the inherent disentanglement capability in stable diffusion can enable powerful image editing that can even outperform the baseline that requires fine-tuning.

Qualitative Comparison with More Baselines: Besides DiffusionCLIP, we have identified three other baselines that use diffusion models for image editing, which are Blended-Diffusion [8], Prompt-to-Prompt [20], and Imagic [34]. However, these methods either have not released code by the time of our submission or require auxiliary labels that are unavailable, so we could not include them in our subjective evaluation. Nevertheless, we managed to generate some qualitative comparisons by collecting the source and edited images shown on their papers and using our method to perform the edit. The edited images are displayed side-by-side in Fig. 6, with each row corresponding to one baseline. As can be observed, our method maintains the semantic content better than Blended Diffusion, which undesirably changes the details such as the grass. Our method performs relatively on par with DiffusionCLIP and Prompt-to-Prompt in terms of both attribute matching and content preservation. Finally, compared with Imagic, our method arguably changes slightly more semantic content (*e.g.,* adding stars) but the overall quality is very competitive. To sum up, these results, and additional results in Appendix F, consistently verify the

Figure 6. Visual comparisons. The source images and the corresponding baseline results are taken from the original papers. Each row indicates the comparison to one particular baseline.

Figure 7. Text descriptions for image editing. The original image is shown in the first column. On the right columns, we demonstrate the results when changing text descriptions $\boldsymbol{c}^{(0)}$ and $\boldsymbol{c}^{(1)}$.

high quality of our edited images against each baseline's most representative results reported in their papers.

4.3. Ablation Study

As described in Sec. 3.3, the optimal $\lambda_{1:T}$ depends on specific values of $\boldsymbol{c}^{(0)}$, $\boldsymbol{c}^{(1)}$, and $\boldsymbol{X}^{T}$. We thus like to investigate whether our image editing algorithm is robust against variations in these inputs. To check the robustness against $\boldsymbol{c}^{(0)}$ and $\boldsymbol{c}^{(1)}$ in the image editing task, we generate different edited images from the same source image by varying the text descriptions. We study three types of variations.

First, we study how the way that $\boldsymbol{c}^{(1)}$ appends the target attribute description can affect the results. Specifically, we fix $\boldsymbol{c}^{(0)}$ and the target attribute, and generate three $\boldsymbol{c}^{(1)}$'s with different ways to concatenate the target attribute: direct concatenation, concatenation separated by a comma, and concatenation separated by "with", as illustrated in the top row of Fig. 7. As can be seen, the resulting edit images are hardly impacted by such variations.

Second, we study whether varying the complexity of the target attribute description in $\boldsymbol{c}^{(1)}$ would affect the results. To this end, we again fix $\boldsymbol{c}^{(0)}$ and gradually add more modifiers describing the same target attribute to $\boldsymbol{c}^{(1)}$, as shown in the second row of Fig. 7. As can be seen, while having more modifiers can amplify the editing effects, having one modifier is sufficient to generate an effective edit.

Finally, we study whether variations in $\boldsymbol{c}^{(0)}$ would affect the results. Specifically, we fix the target attribute description and generate three versions of $\boldsymbol{c}^{(0)}$ that describe the same source object, a short description, a longer description by adding non-informative words, and a description generated by an image captioning model [75]. As shown in the bottom row of Fig. 7, the image editing is largely consistent in almost all cases, but fails with the $\boldsymbol{c}^{(0)}$ generated by the image captioning model. One possible cause is that the generated caption is usually lengthy and contains more details, which may overwhelm the target attribute description.

More image examples are shown in Appendix G, and the overall conclusion is that the edited results are highly robust against most variations in the text descriptions. We also perform some experiments that show strong robustness against variations in the images on which $\lambda_{1:T}$ are learned, which is presented in Appendix H.

5. Conclusion

In this work, we study the disentanglement property in the stable diffusion model. We find that stable diffusion inherently has the disentanglement capability, and it can be triggered by partially replacing the text embedding from a style-neutral description to one with desired style. Motivated by this finding, we propose a simple and light-weight disentanglement algorithm where the combination weights of the two text embeddings are optimized for style matching and content preservation. With only 50 parameters being optimized, our method demonstrates generalizable disentanglement ability and outperforms sophisticated baselines that require fine-tuning on image editing task.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019. 2

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, 2020. 2

[3] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 2018. 2

[4] Johannes Ackermann and Minjun Li. High-resolution image editing via multi-stage blended diffusion. *arXiv preprint arXiv:2210.12965*, 2022. 3, 12, 13

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. 1

[6] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces, 2021. 3

[7] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 3

[8] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 3, 7

[9] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020. 3

[10] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2018. 1, 2

[11] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation, 2020. 3

[12] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Neurips*, 2016. 2

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

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

[15] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *ICLR*, 2018. 2

[16] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[17] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 2022. 4

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Neurips*, 2014. 1

[19] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Neurips*, 2020. 2, 6

[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. 3, 5, 7

[21] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001. 3

[22] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Neurips*, 2020. 1, 3

[24] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *arXiv preprint arXiv:2106.15282*, 2021. 3

[25] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 3

[26] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 3

[27] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation, 2022. 3

[28] Youngjoo Jo and Jongyoul Park. Sc-fegan: Face editing generative adversarial network with user's sketch and color. In *ICCV*, 2019. 3

[29] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 4, 12, 13

[30] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2017. 1

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

[32] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2

[33] 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

[34] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 3, 7

[35] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022. 3, 5, 7, 12
[36] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *ICML*, 2018. 2
[37] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5, 13
[38] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *ArXiv*, abs/2107.00630, 2021. 3
[39] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1
[40] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis, 2020. 3
[41] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 2, 4, 6
[42] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *CVPR*, 2020. 3
[43] Muyang Li, Ji Lin, Chenlin Meng, Stefano Ermon, Song Han, and Jun-Yan Zhu. Efficient spatially sparse inference for conditional gans and diffusion models. *arXiv preprint arXiv:2211.02048*, 2022. 3
[44] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *ICLR*, 2022. 5
[45] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022. 3
[46] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. *arXiv preprint arXiv:2112.05744*, 2021. 3
[47] Yahui Liu, Marco De Nadai, Deng Cai, Huayang Li, Xavier Alameda-Pineda, Nicu Sebe, and Bruno Lepri. Describe what to change: A text-guided unsupervised image-to-image translation approach. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 3
[48] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 7
[49] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 3
[50] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3, 5
[51] 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, 2021. 3
[52] Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, Philip Torr, et al. Learning disentangled representations with semi-supervised deep generative models. In *Neurips*, 2017. 2
[53] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 2, 3, 4
[54] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 2022. 3
[55] 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 *ICML*, 2021. 3
[56] 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. 3
[57] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Neurips*, 2019. 1
[58] Xuanchi Ren, Tao Yang, Yuwang Wang, and Wenjun Zeng. Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view. In *International Conference on Learning Representations*, 2021. 2, 12, 13
[59] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, 2015. 1
[60] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014. 1
[61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 5
[62] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 2
[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. *arXiv preprint arXiv:2208.12242*, 2022. 3
[64] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 3
[65] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3
[66] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, pages 1–14, 2022. 3

[67] 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

[68] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 2, 3

[69] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2, 3, 6

[70] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. 2

[71] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2015. 1, 3, 12

[72] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 3

[73] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 1, 3

[74] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation, 2021. 3

[75] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022. 8, 18

[76] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 7, 12

[77] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2017. 3