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

Qiucheng Wu\textsuperscript{1}, Yujian Liu\textsuperscript{1}, Handong Zhao\textsuperscript{2}, Ajinkya Kale\textsuperscript{2}, Trung Bui\textsuperscript{2}, Tong Yu\textsuperscript{2}, Zhe Lin\textsuperscript{2}, Yang Zhang\textsuperscript{3}, Shiyu Chang\textsuperscript{1}

\textsuperscript{1}UC, Santa Barbara, \textsuperscript{2}Adobe Research, \textsuperscript{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-
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. Par-

<table>
<thead>
<tr>
<th>Scenes</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Global</strong></td>
<td><strong>Local</strong></td>
</tr>
<tr>
<td>Styles (children drawing, cyberpunk, anime), Building appearance</td>
<td>Cherry blossom, rainbow, foothills</td>
</tr>
<tr>
<td>(wooden, red brick), Weather &amp; time (sunset, night, snowy)</td>
<td></td>
</tr>
<tr>
<td><strong>Expressional</strong></td>
<td><strong>Small edits</strong></td>
</tr>
<tr>
<td>Styles (renaissance, Egyptian mural, sketch, Pixar)</td>
<td>Cake toppings, remove people on the street</td>
</tr>
<tr>
<td>Appearance (young, tanned, male)</td>
<td>Hats, hair colors, earrings</td>
</tr>
</tbody>
</table>

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.
ticularly, 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, 55, 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 $c$, the goal of the text-conditioned DDIM is to generate an image, denoted as $X_0$, that conforms to the text description. DDIM defines a set of noisy images, $X_{1:T}$, by adding Gaussian noise to $X_0$ according to a predefined diffusion process. Each $X_t$ is corrupted with a larger noise than $X_{t-1}$, and $X_T$ is very close to standard Gaussian noise. The generation process of DDIM tries to denoise from $X_T$ all the way back to $X_0$. Specifically, in the first step, $X_2$ is randomly drawn from a standard Gaussian distribution. Then, each $X_{t-1}$ is inferred from $X_t$ via the following denoising process:

$$X_{t-1} = \gamma_{t0} X_t + \gamma_{t1} \epsilon_t(X_t, t, c_t),$$

where $c_t$ is the text embedding used at step $t$. In most common settings, generating one image only requires one text description $c$, so $c_t = c$ for all $t$. Here we make it dependent upon $t$ to accommodate the discussions in the following subsections. $\epsilon_t(X_t, c_t)$ is a pre-trained denoising network that infers $X_0$ given the input of $X_t$ and $c_t$. The parameters of the denoising network, $\theta^*$, are considered fixed throughout this section. $\gamma_{t0}$ and $\gamma_{t1}$ are defined as

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

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 $X_T$ and the text descriptions $c_{1:T}$. Thus we introduce the following notation,

$$X_0 = g(X_T, c_{1:T}),$$

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.
entangled generation of image using consistent, as shown in Appendix. Performments on more objects and styles and the observations are with full replacement. We have performed the same exper-

clement capability in stable diffusion models, which can be triggered by partially replacing the text embeddings, but not with full replacement. 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, \( c^{(0)} \) and \( c^{(1)} \). \( c^{(0)} \) is the embedding of a style-neutral description, “a photo of person”; and \( c^{(1)} \) is the embedding of a description with an explicit style, “a photo of person with smile”. When \( c^{(0)} \) is fed to the model, the generated image is a person with neutral explicit style, “a photo of person with smile”. When \( c^{(0)} \) is fed to the model, the generated image is a person with neutral explicit style, “a photo of person with smile”. When \( c^{(0)} \) is fed to the model, the generated image is a person with neutral explicit style, “a photo of person with smile”. When \( c^{(0)} \) is fed to the model, the generated image is a person with neutral explicit style, “a photo of person with smile”. When \( c^{(0)} \) is fed to the model, the generated image is a person with neutral explicit style, “a photo of person with smile”.

Case 1: Full Replacement. In our first attempt, we re-

duced, \( c^{(1)} \), is the embedding of a style-neutral description, “a photo of person”; and \( c^{(1)} \) is the embedding of a description with an explicit style, “a photo of person with smile”. When \( c^{(0)} \) is fed to the model, the generated image is a person with neutral explicit style, “a photo of person with smile”. When \( c^{(0)} \) is fed to the model, the generated image is a person with neutral explicit style, “a photo of person with smile”. When \( c^{(0)} \) is fed to the model, the generated image is a person with neutral explicit style, “a photo of person with smile”.

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

Our goal is to find an optimal \( \lambda_{1:7} \) such that \( X^{(0)} \) maintains the same semantic content as \( X^{(0)} \) but conforms to the style described in \( c^{(1)} \), which is achieved by solving the following optimization problem similar to [41, 53]:

\[
\min_{\lambda_{1:7}} \mathcal{L}_{\text{clip}}(X^{(0)}, X^{(0)}, c^{(0)}, c^{(1)}) + \beta \mathcal{L}_{\text{perc}}(X^{(0)}, X^{(0)}).
\]

where \( \lambda_{i} \) 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 \( \lambda_{i} \) to either 1 or 0).

Given a random noise \( X_{T} \) and the text description pair \( c^{(0)} \) and \( c^{(1)} \), the optimization procedure for \( \lambda_{1:7} \) is as follows. First, two images are generated, a style-neutral one generated with \( c^{(0)} \) and the other generated with \( c^{(1)} \), namely

\[
X^{(0)} = g(X_{T}, c^{(0)}), \quad X^{(0)} = g(X_{T}, c^{(1)}).
\]

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 different 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 \( c^{(0)} \), the neutral image

\[
c_{t} = \lambda_{t} c^{(1)} + (1 - \lambda_{t}) c^{(0)} \equiv c^{(j)},
\]

where \( \lambda_{t} \) 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 \( \lambda_{i} \) to either 1 or 0).

Given a random noise \( X_{T} \) and the text description pair \( c^{(0)} \) and \( c^{(1)} \), the optimization procedure for \( \lambda_{1:7} \) is as follows. First, two images are generated, a style-neutral one generated with \( c^{(0)} \) and the other generated with \( c^{(1)} \), namely

\[
X^{(0)} = g(X_{T}, c^{(0)}), \quad X^{(0)} = g(X_{T}, c^{(1)}).
\]

Our goal is to find an optimal \( \lambda_{1:7} \) such that \( X^{(0)} \) maintains the same semantic content as \( X^{(0)} \) but conforms to the style described in \( c^{(1)} \), which is achieved by solving the following optimization problem similar to [41, 53]:

\[
\min_{\lambda_{1:7}} \mathcal{L}_{\text{clip}}(X^{(0)}, X^{(0)}, c^{(0)}, c^{(1)}) + \beta \mathcal{L}_{\text{perc}}(X^{(0)}, X^{(0)}).
\]

where \( \lambda_{i} \) 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 \( \lambda_{i} \) to either 1 or 0).

Given a random noise \( X_{T} \) and the text description pair \( c^{(0)} \) and \( c^{(1)} \), the optimization procedure for \( \lambda_{1:7} \) is as follows. First, two images are generated, a style-neutral one generated with \( c^{(0)} \) and the other generated with \( c^{(1)} \), namely

\[
X^{(0)} = g(X_{T}, c^{(0)}), \quad X^{(0)} = g(X_{T}, c^{(1)}).
\]

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

\[
\mathcal{L}_{\text{perc}}(X^{(0)}, X^{(0)}) = \| h(X^{(0)}) - h(X^{(0)}) \|_{1},
\]

where \( h(\cdot) \) 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 \( \lambda_{1:7} \), so this optimization problem involves very few parameters and does not need to fine-tune the diffusion model itself.
Again, for notational brevity, we summarize this new generation process as
\[ \bar{X}_0 = \hat{g}(\bar{X}_T, c_1, r_1^T). \] (13)
It can be easily shown that \( \hat{g}(\bar{X}_T, c^{(0)}, E_1, r_1^T) = I. \)

Figure 3. Overview of our method that finds optimal text embedding for disentanglement. In this example, \( c^{(0)} \) is the embedding of “A castle”, and \( c^{(1)} \) is the embedding of “A children drawing of castle”. First two rows: optimization process that finds the best soft combination of \( c^{(0)} \) and \( c^{(1)} \), 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 \( I \) to show the distinction. Therefore, if we could find a value for initial random variable \( X_T \) such that the stable diffusion model can generate exactly the same image as \( I \) when conditioned on \( c^{(0)} \), 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, \( \bar{X}_{1:T} \) based on \( I \) as follows:
\[ \bar{X}_0 = I, \quad \bar{X}_{t+1} = \gamma_0^{(t)} \bar{X}_t + \gamma_1^{(t)} e_\theta(\bar{X}_t, t, c^{(0)}), \] (9)
where
\[ \gamma_0 = \sqrt{1 - \alpha_t^{-1}}, \quad \gamma_1 = \sqrt{1 - \alpha_t^{-1} - \alpha_{t+1}^{-1}}. \] (10)

It has been shown [35] that by setting \( X_T = \bar{X}_T \) and following through the generation process in Eq. (1) with \( e_t = c^{(0)} \), the resulting \( X_{0:T-1} \) would satisfy \( X_t \approx \bar{X}_t, \forall t \in [0, T-1] \), and therefore \( X_0 \approx I \).

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 \( \bar{X}_T = X_T \), and then
\[ \bar{X}_{t-1} = \gamma_0^{(t)} \bar{X}_t + \gamma_1^{(t)} e_\theta(\bar{X}_t, t, c_t) + E_t, \] (11)
where the \( E_t \) is the correction term defined as
\[ E_t = \hat{X}_{t-1} - \gamma_0^{(t)} \bar{X}_t - \gamma_1^{(t)} e_\theta(\bar{X}_t, t, c^{(0)}). \] (12)

Again, for notational brevity, we summarize this new generation process as
\[ \bar{X}_0 = \hat{g}(\bar{X}_T, c^{(0)}, E_1, r_1^T). \] (13)
Now that we have developed a generation process that can reconstruct \( I \), we can now follow the same procedure in Sec. 3.3 to perform image editing, with the image generation in Eq. (6) replaced with
\[ \bar{X}_0^{(0)} = \hat{g}(\bar{X}_T, c^{(0)}, E_1, r_1^T) = I, \quad \bar{X}_0^{(\lambda)} = \hat{g}(\bar{X}_T, c^{(\lambda)}, E_1, r_1^T). \] (14)

It is worth emphasizing that when generating \( \bar{X}_0^{(\lambda)} \), the error correction terms \( E_1, r_1^T \) are still fixed to the ones computed for reconstructing \( I \). 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 pretrained 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 \( \lambda_1, \tau \), we use Adam [37] optimizer with learning rate 0.03. To balance two loss terms, we set \( \beta \) 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...
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 \( c^{(0)} \) and the description with style \( c^{(1)} \). In particular, \( c^{(0)} \) includes just the name of the object, e.g., “A castle”, and \( c^{(1)} \) is constructed by appending the description of the attribute to \( c^{(0)} \), separated by comma (e.g., “A castle, children drawing style”). Next, we generate five style-neutral images for each \( c^{(0)} \) 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
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 $\lambda_{1,7}$ 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 CelebA [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 BLENDDED-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 BLENDDED-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
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 $c^{(0)}$, $c^{(1)}$, and $X^{T}$. We thus like to investigate whether our image editing algorithm is robust against variations in these inputs. To check the robustness against $c^{(0)}$ and $c^{(1)}$ 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 $c^{(1)}$ appends the target attribute description can affect the results. Specifically, we fix $c^{(0)}$ and the target attribute, and generate three $c^{(1)}$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 $c^{(1)}$ would affect the results. To this end, we again fix $c^{(0)}$ and gradually add more modifiers describing the same target attribute to $c^{(1)}$, 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 $c^{(0)}$ would affect the results. Specifically, we fix the target attribute description and generate three versions of $c^{(0)}$ 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 $c^{(0)}$ 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


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


and editing with text-guided diffusion models, 2021.


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


