COMPOSITIONAL IMAGE GENERATION AND MANIPULATION WITH DIFFUSION PROBABILISTIC MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a principled method for compositional image generation and manipulation using diffusion probabilistic models. In particular, for any pre-trained generative model with a semantic latent space, we train a latent diffusion model and auxiliary latent classifiers to help navigate latent representations in a non-linear fashion. We show that such conditional generation achieved by latent classifier guidance provably maximizes a lower bound of the conditional log-likelihood during training, and can reduce to a simple latent arithmetic method with additional assumption, which is surprisingly under-studied in the context of compositionality. We then derive a new guidance term which is shown to be crucial for maintaining the original semantics when doing manipulation. Unlike previous methods, our method is agnostic to pre-trained generative models and latent spaces, while still achieving competitive performance on compositional image generation as well as sequential manipulation of real and synthetic images.

1 INTRODUCTION

In recent years, the machine learning and computer vision communities have witnessed great progress in the field of deep generative modeling. From variational autoencoders (VAEs) (Kingma & Welling, 2013), normalizing flows (Rezende & Mohamed, 2015), and generative adversarial networks (GANs) (Goodfellow et al., 2014; Brock et al., 2018; Choi et al., 2020; Abdal et al., 2020; Wu et al., 2021), to the very recent diffusion probabilistic models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020a; Nichol & Dhariwal, 2021; Abstreiter et al., 2021; Rombach et al., 2022) and score-based models (Song & Ermon, 2019; Song et al., 2020b), generating high-quality, realistic images has become easier, if not impossible before. Despite the significant progress that has been made, controlling the generation process using various conditions, such as class labels and text descriptions, still remains challenging.

One major difficulty towards such controllable generation is compositionality. Compositionality in generative modeling, or compositional generation, is the ability of a conditional generative model to produce realistic outputs given multiple conditions and their relations. Broadly speaking, there exist two types of methods for achieving such compositionality. The first type of methods focus on the latent space of pre-trained generative models. Their goal is a rule that governs the manipulation of latent codes so as to obtain outputs with desired properties. When the latent space is disentangled as in StyleGAN (Karras et al., 2019; 2020), linear control is possible by carefully identifying and combining the latent directions of each attribute (Shen et al., 2020; Wu et al., 2021; Härkönen et al., 2020; Shen & Zhou, 2021). While this is not new, the feasibility of such linear control in the context of compositionality is still under-explored. On the other hand, non-linear manipulations of the latent space have also been proposed for finer control, in the sense that each modification will be customized for each latent code. However, the previous non-linear methods are either not amendable for new attributes (Abdal et al., 2021) or not agnostic to various models and latent spaces (Nie et al., 2021). The second type of methods tackle the problem in the image space, either relying on energy-based models (EBMs) or borrowing ideas from them (Du et al., 2020; Liu et al., 2022). However, such methods can not leverage the nice properties of the latent space such as disentanglement (Bengio et al., 2013), and training multiple sub-models in image space can be cumbersome.

In this paper, we propose a principled method for compositional image generation and manipulation[1] to alleviate previously mentioned problems. Our method leverages the latest advances with diffusion probabilistic models to manipulate latent spaces in a non-linear way, but also allows a simplified linear version resembling vector arithmetic-based manipulation. For compositional generation, we train latent diffusion models and auxiliary latent classifiers for pre-trained generators, and use classifier guidance to sample in the latent space. For manipulating given synthetic or real images, an additional guidance term is induced by formulating the problem as adding a source image condition to compositional generation. We show that such conditioning on multiple attributes achieved by latent classifier guidance provably maximizes a lower bound of the conditional log-likelihood, and reduces to simple linear vector arithmetic under additional constraints. Experiments show that our method can generate realistic images with different attribute compositions, and can manipulate attributes of both **synthetic and real** images in a coherent way. We also demonstrate that our linear version based on vector arithmetic can serve as a strong baseline in many scenarios, despite of previous non-linear manipulation studies. Notably, unlike the previous method (Nie et al., 2021), we train DDIM (Song et al., 2020a) to model latent space thus our method is latent-space-agnostic and thus can be applied to any latent-variable generative models, meaning it works for different pre-trained generative models with a compact latent space.

In summary, our contributions are as follows:

1. We leverage recent progress in conditional diffusion probabilistic models to achieve compositional generation and manipulation under a unified perspective, and connect the proposed framework to vector arithmetic in latent spaces.
2. We show that our method is both model-agnostic and space-agnostic. This enables manipulation on new models such as (Preechakul et al., 2022) and expanded spaces such as (Abdal et al., 2019), which is crucial for obtaining high-quality real image editing.
3. With state-of-the-art pre-trained generative models, we discuss the questions of why and when linear manipulation can perform reasonably well, and the scenarios when non-linear methods should be considered.

2 METHOD

In this section, we will describe how we leverage latent diffusion models and classifier guidance to generate and manipulate images in a principled way.

2.1 LATENT DIFFUSION MODELING

A diffusion model is a deep latent variable model that maps a pre-defined noise distribution to data distribution through smooth, iterative denoising steps. Formally, given a unknown data distribution $p(x)$, a diffusion model approximates the distribution with the following form $p_\theta(x_0) = \int p_\theta(x_{0:T})dx_{1:T}$, where $x_{1:T}$ are the latent variables of the same dimensionality as the data x_0. The forward diffusion process, resembling a parameter-free encoder, is a Markov chain $q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$, where each $q(x_t|x_{t-1})$ is typically a Gaussian distribution. The forward process perturbs inputs according to a pre-defined schedule, and the transformed data distribution $q(x_t|x_0)$ will gradually converge to a standard Gaussian $\mathcal{N}(x_T; 0, I)$. The reverse process, resembling a hierarchical decoder, is composed of a sequence of de-noising steps $p_\theta(x_{t-1}|x_t)$, which is parameterized by a deep neural network with parameter θ.

During training, the input images are corrupted by the forward process, and the diffusion model is trained to convert the corrupted images back to the original inputs. For de-noising diffusion probabilistic models (DDPM) (Ho et al., 2020), the training objective is derived as a re-weighted variational bound by treating DDPMs as VAEs, and for scored-based generative models (Song & Ermon, 2019), the objective is derived using score matching. To generate from the learned distribution, one first samples x_T from a standard Gaussian and uses the reverse process to transform it into the image space.

[1]For the sake of clarity, the term "compositional generation" will refer to generating images conditioning only on attributes while the term "compositional manipulation" will specifically refer to conditioning on both attributes and an original image.

Here, we focus on the latent space of a pre-trained generative model. For a pre-trained generator G that maps a latent space $\mathcal{Z}$ to the image space $\mathcal{X}$, we train a diffusion model to approximate the latent distribution $p(z)$. One of the advantages of modeling the latent space rather than the image space is that the latent space enjoys some properties that image space does not, such as disentanglement (Bengio et al., 2013). Another advantage is the ability to use various guidance techniques in the latent space, since training latent guidance terms is typically easier than training other manipulation methods in image space (Shen et al., 2020).

2.2 Conditional and Compositional Generation

Conditional generation with diffusion models relies on perturbing unconditional generation with user-specified guidance terms, namely classifier guidance (Sohl-Dickstein et al., 2015; Song et al., 2020b; Dhariwal & Nichol, 2021) and classifier-free guidance (Ho & Salimans, 2022). Although classifier-free guidance performs competitively in image space and is sometimes more favorable than classifier guidance (Ho & Salimans, 2022; Liu et al., 2022), here we lean on classifier guidance in our latent diffusion models for several reasons. We argue that the cause of classifier guidance falling behind classifier-free guidance in the image space is that image classifiers tend to learn shortcuts from suspicious correlations, while this problem is alleviated in a compact, even disentangled latent space. Another benefit of using classifier guidance is that classifiers are usually easier to train than classifier-free guidance. Finally, when the classifiers are linear, classifier guidance resembles linear arithmetic methods, as we will show in Section 2.4.

The goal of conditional generation is to model the conditional distribution $p(z|y)$ where y is the conditions or attributes. By Bayes rules $p(z_t|y) = {}^{p(z_t)p(y|z_t)}/_{p(y)}$, the score of the conditional probability $\nabla_{z_t} \log p(z_t|y)$ can be factorized as the unconditional score $\nabla_{z_t} \log p(z_t)$ and the gradient flow $\nabla_{z_t} \log p(y|z_t)$. Therefore, one simply needs an unconditional latent diffusion model and a latent classifier to model the conditional score, known as classifier guidance. In practice, the classifier guidance term is usually scaled by a factor α, such that $\nabla_{z_t} \log p(z_t|y) = \nabla_{z_t} \log p(z_t) + \alpha \nabla_{z_t} \log p(y|z_t)$. The factor α serves as a temperature parameter which adds another layer of controllability to the sharpness of the posterior distribution $p(y|z_t)$.

Compositional generation can be considered as conditional generation with multiple conditions and the relations among them. In this paper, we consider two relations, conjunction "AND" and negation "NOT". For the conjunction of attributes $y^1 \wedge y^2 \wedge ... \wedge y^n$, assuming the conditions to be independent of each other, we can simply factorize the "compositional" log-probability as

$$\nabla_{z_t} \log p(z_t|y^1, y^2, ..., y^n) = \nabla_{z_t} \log p(z_t) + \sum_{i=1}^{n} \alpha_t^i \nabla_{z_t} \log p(y^i|z_t). \tag{1}$$

And with attribute negations $y^1 \wedge ... \wedge y^{m-1} \wedge \overline{y^m} \wedge ... \wedge \overline{y^n}$, without loss of generality, we can factorize the log-probability similarly

$$\nabla_{z_t} \log p(z_t|y^1, ..., y^n) = \nabla_{z_t} \log p(z_t) + \sum_{i=1}^{m-1} \alpha_t^i \nabla_{z_t} \log p(y^i|z_t) - \sum_{i=m}^{n} \beta_t^i \nabla_{z_t} \log p(y^i|z_t). \tag{2}$$

While classifier guidance is useful for compositional generation, there is no guarantee that those results will be anything similar to the original image when doing manipulations. This is because the generation is not conditioned on the original image. As there is no constraint on the specific form of the posterior (Sohl-Dickstein et al., 2015), conditioning on the original image amounts to adding a new guidance term $\gamma_t \nabla_z \log p(\hat{z}|z)$, where $\hat{z}$ is the latent of the image to be manipulated. For conjunction relations as in Equation (1), the overall score function for manipulation then becomes

$$\nabla_{z_t} \log p(z_t|y^1, y^2, ..., y^n, \hat{z}) = \nabla_{z_t} \log p(z_t) + \sum_{i=1}^{n} \alpha_t^i \nabla_{z_t} \log p(y^i|z_t) + \gamma_t \nabla_{z_t} \log p(\hat{z}|z_t), \tag{3}$$

similarly for Equation (2) with the presence of negation. When $p(\hat{z}|z_t)$ is modeled by a Gaussian distribution, the new guidance term behaves as a regularization term $\nabla_{z_t} \|z_t - \hat{z}\|_2^2$.

2.3 MODEL TRAINING

We argue that a true compositional model should be able to easily include new attributes without re-training the whole model. In our method, the training of the unconditional diffusion models and the latent classifiers can be decoupled, and such training amounts to maximizing the evidence lower bound (ELBO) of the conditional log-likelihood. This means that whenever the model is given new attributes, our method simply requires training classifiers on these new attributes, and the latent diffusion model as well as used classifiers can be recycled.

We take DDPM as our example and begin with unconditional generation.

Lemma 1. *The unconditional ELBO of DDPM is given by the following equation*

$$\mathcal{L}_{uncond} := \mathbb{E}_{q(z_{1:T}|z_0)}\left[\log \frac{p(z_T)}{q(z_T|z_0)} + \sum_{t=2}^{T} \log \frac{p(z_{t-1}|z_t)}{q(z_{t-1}|z_t, z_0)} + \log p(z_0|z_1)\right]. \tag{4}$$

See Ho et al. (2020) for the detailed proof.

Lemma 2 (Compositional generation and manipulation)**.** *The conditional ELBO of DDPM with condition y is given by*

$$\mathcal{L}_{uncond} + \mathbb{E}_{q(z_{1:T}|z_0)}\left[\sum_{t=1}^{T} \log p(y|z_{t-1})\right] + C, \tag{5}$$

and with independent conditions $\{y^1, y^2, ..., y^n\}$ and $\hat{z}$

$$\mathcal{L}_{uncond} + \mathbb{E}_{q(z_{1:T}|x_0)}\left[\sum_{t=1}^{T}\left[\sum_{i=1}^{n} \log p(y^i|z_{t-1}) + \log p(\hat{z}|z_{t-1})\right]\right] + C. \tag{6}$$

The full derivation is shown in Section A.1. Lemma. 2 states that training unconditional diffusion models and their latent classifiers is equivalent to maximizing the ELBO of joint log-likelihood of z and y up to a constant.

2.4 CONNECTION TO LINEAR ARITHMETIC

Our regularized method manipulates a given latent $\hat{z}$ in a non-linear fashion, but it degrades to linear manipulation with additional assumptions. We take the case where there are only conjunction relations as an example and consider Equation (3).

Lemma 3 (Compositional manipulation and linear arithmetic)**.** *When $p(z_t)$ is non-informative and $\log p(y|z_t)$ are linear, the proposed manipulation is endowed with an analytic solution*

$$z_0 = \hat{z} + \frac{1}{\gamma_0}\sum_{i=1}^{n} \alpha_0^i w^i. \tag{7}$$

Proof. We first assume that $p(z_t)$ is a non-informative distribution where $\nabla_{z_t} p(z_t) = 0$. Then we model each $\log p(y^i|z_t)$ with a linear classifier $z \mapsto w^T z + b$, so that the gradient $\nabla_{z_t} \log p(y^i|z_t) \sim w$ up to a scale factor[2]. Now when the reverse process of latent diffusion model converges at $t = 0$, the whole Equation (3) should converge to 0

$$\sum_{i=1}^{n} \alpha_0^i w^i + \gamma_0(z_0 - \hat{z}) = 0, \tag{8}$$

which leads to the above analytic solution. □

For attribute negation, the solution perturbs $\hat{z}$ towards the negative direction of the classifiers. This is a natural multi-attributes generalization of the vector arithmetic method, and we refer it as the linear version of our method in later comparisons.

[2]Let the scalar absorbed by α_t^i.

3 Related Work

(Conditional) diffusion models Diffusion models have become increasingly favored over other generative models such as GANs and VAEs due to their photo-realistic generation quality and ease of training. Sohl-Dickstein et al. (2015) proposed the first working concept of diffusion models from the perspective of thermodynamics, then this concept was followed by Song & Ermon (2019); Song et al. (2020b); Ho et al. (2020); Song et al. (2020a) whose works established the foundation of diffusion models that we see today. For the conditional generation with diffusion models, Dhariwal & Nichol (2021) further formulated classifier guidance and lifted the generation quality of diffusion models over previous state-of-the-art GANs. Ho & Salimans (2022) then proposed classifier-free guidance which nowadays is used in many large-scale image generation engine (Nichol et al., 2021; Ramesh et al., 2022; Saharia et al., 2022). Although most of these diffusion models work on the image space, recently latent diffusion models have also drawn great attention and achieved remarkable results (Vahdat et al., 2021; Abstreiter et al., 2021; Rombach et al., 2022).

Compositionality in latent space Due to the wide success of StyleGANs (Karras et al., 2019; 2020) in image generation, most efforts on conditional generation have been focusing on the latent space of StyleGANs. To use linear arithmetic for manipulation, various ways of finding attribute directions have been proposed. Shen et al. (2020) find a direction for each attribute by training linear SVMs, then perturb the latent point along the orthogonal projection of these directions to prevent unwanted semantic changes. Wu et al. (2021) detect latent channels that only allow local changes for specific attributes. Such directions can also be identified in an unsupervised fashion, using the PCA decomposition of the latent space Härkönen et al. (2020), or the SVD of the first subsequent linear layer Shen & Zhou (2021). On the non-linear side, (Abdal et al., 2021) design a framework to control a set of pre-decided attributes using conditional normalizing flow. (Nie et al., 2021) use latent EBMs to control the style generation with non-linear classifiers. Note that none of the linear methods explicitly discussed the composition of multiple attributes and their relations as in the non-linear methods.

Compositionality in image space Some other methods tackle compositional generation in the image space. These methods either directly use EBMs or use diffusion models that can be considered as EBMs. Du et al. (2020) trains EBMs for each condition and composes them together by defining new energy functions based on each individual energy function and their relations. Liu et al. (2022) followed their proposal and adopted classifier-free guidance in diffusion models to multiple conditions. However, these image space models can not leverage the disentanglement property of the latent space, and training EBMs or diffusion models for each condition can be cumbersome. Note that although some large-scale text-to-image generation engines (Nichol et al., 2021; Ramesh et al., 2022; Saharia et al., 2022) claim compositionality in their methods, they do not model compositionality explicitly but rather rely on implicit composition by their language models, which leads to less satisfying results when the set of conditions gets large (Liu et al., 2022).

4 Experiments

We evaluate our method on two pre-trained models, StyleGAN2 (Karras et al., 2020) and Diffusion Autoencoder (Preechakul et al., 2022). To use our methods in the intermediate latent space ($\mathcal{W}_{\text{s}}$ space) of a pre-trained StyleGAN2, we first train latent DDIM (Song et al., 2020a) on 100,000 w_{s} vectors sampled from the pushforward distribution given by the style generation. We then train linear classifiers on the $\mathcal{W}_{\text{s}}$ space using the latent-label pairs provided by Abdal et al. (2021). Note that although Equation (5) requires classifiers to be time-dependent, we find that using the same linear classifiers trained on clean w_{s} vectors can still produce reasonable results. Details on the architecture of the latent diffusion models and the performance of the classifiers can be found in Appendix A.2 and Appendix A.3. For Diffusion Autoencoder, we use their pre-trained latent diffusion model and linear classifiers.

For real image manipulation, we also need to encode input images to the latent space. With StyleGAN2, we use the optimization-based inversion method in (Karras et al., 2020) to get the initial latent space $\mathcal{Z}_{\text{s}}$ and intermediate $\mathcal{W}_{\text{s}}$ space encodings and use the pre-trained pSp encoder (Richardson et al., 2021) to get $\mathcal{W}_{\text{s}}+$ space encodings, where $\mathcal{W}_{\text{s}}+$ is a concatenation of 18 different 512-

(a) LACE-LD (Nie et al., 2021) (b) LACE-ODE (Nie et al., 2021)

(c) Ours-Linear (d) Ours-Diffusion

Figure 1: Visual comparison on 1024×1024 compositional generation using StyleGAN2 generator. The target attribute conditions are: female, smile, and 55 y/o. "y/o" means "years old".

(a) Ours-Linear (b) Ours-Diffusion

Figure 2: Visual comparison on 256×256 compositional generation using Diffusion Autoencoder generator. The target attribute conditions are: male, smile, and glasses.

dimensional w_{s} vectors in StyleGAN2. With Diffusion Autoencoder, we can directly use their pre-trained encoders to get semantic vectors.

Following Nie et al. (2021), we consider three metrics for our evaluation: Fréchet Inception Distance (FID) (Heusel et al., 2017), face identity loss (ID) (Abdal et al., 2021) and conditional accuracy (ACC). FID measures generation quality by comparing the Inception feature distribution of generated outputs and real images. ID reflects the ability of a manipulation method to preserve the identity of an input face. It uses a pre-trained face recognition model to generate embeddings of a pair of input and manipulated face images and computes the distance between them. ACC measures the efficacy of manipulation, which is the accuracy of classifying attributes of generated images with randomly sampled target conditions using off-the-shelf image classifiers.

4.1 Compositional Generation

We first evaluate the ability of our methods to generate images with multiple desired attributes. For high-resolution images (1024×1024), we select StyleGAN2 as our pre-trained generator; for low-resolution images (256×256), we use Diffusion Autoencoder.

We compare our method with StyleFlow (Abdal et al., 2021) and LACE (Nie et al., 2021). Numbers of StyleFlow are directly taken from (Nie et al., 2021) for the conjunction of "gender", "smile" and "age", while it cannot handle the other compositional task where negation relations are involved. To compare with LACE, we use their official implementation. Note that LACE is not applicable for Diffusion Autoencoder as its semantic latent space is not endowed with a parameterized distribution

Table 1: StyleGAN2 Compositional Generation

Method	gender, smile, age				-gender, smile, -haircolor			
	FID ↓	ACC ↑			FID ↓	ACC ↑		
		gender	smile	age		gender	smile	haircolor
StyleFlow (Abdal et al., 2021)	43.88	0.718	0.870	0.874	—	—	—	—
LACE-LD (Nie et al., 2021)	22.34	0.953	0.954	0.925	22.86	0.678	0.958	0.924
LACE-ODE (Nie et al., 2021)	22.03	0.964	0.967	0.925	23.51	0.649	0.970	0.935
Ours-Linear	22.46	0.980	0.982	0.863	23.94	0.948	0.995	0.936
Ours-Diffusion	26.49	0.981	0.968	0.863	29.62	0.987	0.954	0.906

Table 2: Sequential Editing with StyleGAN2

Method	FID ↓	ID ↓	ACC ↑			
			yaw	smile	age	glasses
StyleFlow (Abdal et al., 2021)	44.13	0.549	0.947	0.773	0.817	0.876
LACE-ODE (Nie et al., 2021)	27.49	0.501	0.938	0.956	0.881	0.997
Ours-Linear	29.48	0.290	0.887	0.983	0.875	0.786
Ours-Diffusion	24.06	0.445	0.903	0.963	0.845	0.843

as the $\mathcal{Z}_s$ or $\mathcal{W}_s$ space of StyleGAN2. However, our method which fits any latent distributions with diffusion models can be applied.

The quantitative comparison is shown in Table 1 and the qualitative comparison shown in Figure 1 and Figure 2. While for the quantitative results target conditions are randomly sampled for each attribute, for the qualitative results, we use fixed targets for the sake of visualization. As we can see, both versions of our methods, based on linear arithmetic and latent diffusion models, perform competitively against previous non-linear methods.

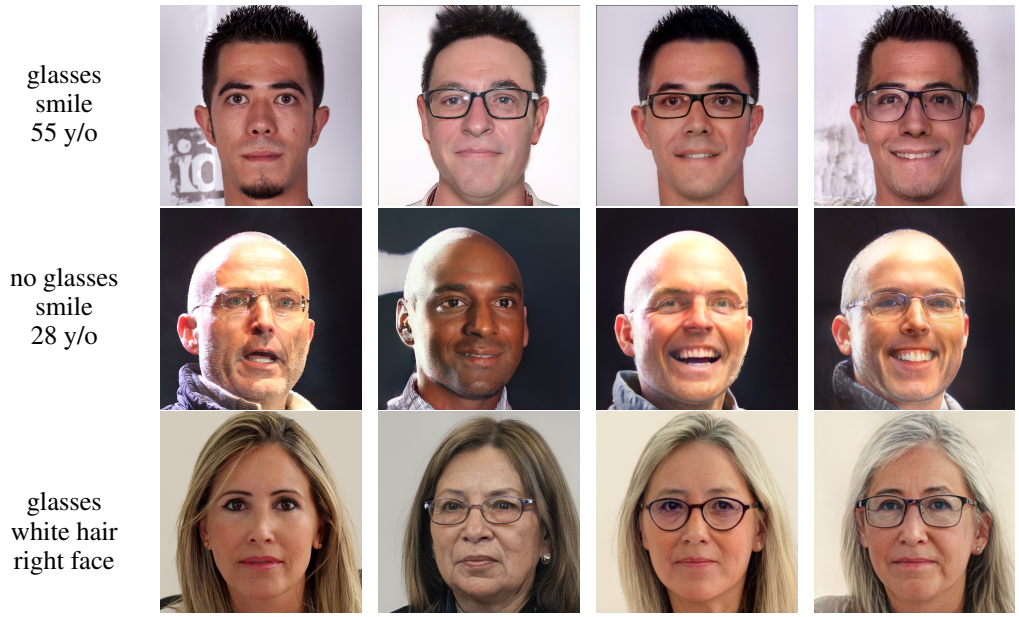
4.2 Compositional Manipulation

We evaluate our methods on manipulating both synthetic and real images.

Synthetic Images To evaluate synthetic image manipulation, we first sample latent codes in $\mathcal{Z}_s$ space and $\mathcal{W}_s$ space, then generate their corresponding output images. To ensure fair comparisons, we use the style network of StyleGAN2 to generate w_s vectors following (Nie et al., 2021) rather than sample vectors from the latent diffusion model that we learn. We then sequentially edit each synthetic image given the target conditions. Results are shown in Table 2. Note that both our linear arithmetic based and latent diffusion model based method achieves competitive FID and ID scores with most attributes successfully manipulated (except for "glasses").

Real Images Manipulating real images can be much harder than manipulating synthetic images as it sometimes involves inverting source images to their latent codes. Our method being space-agnostic brings additional advantages when editing real images. It is well-known that not all real images can be encoded into the $\mathcal{Z}_s$ space and $\mathcal{W}_s$ space of StyleGAN, and expanded spaces such as $\mathcal{W}_s$+ space (Abdal et al., 2019) and $\mathcal{S}$ space (Wu et al., 2021) are better choices for real image editing. However, LACE is restricted to the intermediate space of StyleGAN and thus cannot leverage the richness of the expanded spaces. Moreover, it also requires inverting input images to $\mathcal{Z}_s$ space, which is very challenging. Our method, on the other hand, can be easily adapted to existing or new expanded spaces, where the semantics are richer and the inversion is easier.

As shown in Figure 3, our method outperforms LACE in terms of real image editing. For LACE, the identities of the manipulation change dramatically in all three cases. This is because it is generally hard to invert real input images to $\mathcal{Z}_s$ space, which is required for LACE's manipulation. On the other hand, our methods which only require inversion into $\mathcal{W}_s$ or $\mathcal{W}_s+$ space can control attributes better as well as preserves the identity more faithfully than LACE. $\mathcal{W}_s$ space manipulation controls the attributes very well, but the image quality is sub-optimal due to the limited expressiveness of $\mathcal{W}_s$ space. $\mathcal{W}_s+$ space manipulation provides better image quality, but the attributes are harder



(a) Original (b) LACE-ODE (c) Ours-Linear-$\mathcal{W}_s$ (d) Ours-Linear-$\mathcal{W}_s$+

Figure 3: Visual comparison between different compositional manipulation methods on real image inputs. Target attributes are shown on the left. "y/o" means "years old".

to control, e.g. the "glasses" attribute in the second row. This is because $\mathcal{W}_s+$ space has higher dimensions and training well-behaved classifiers can be harder due to problems such as over-fitting.

5 Discussion

5.1 Why is the linear method competitive?

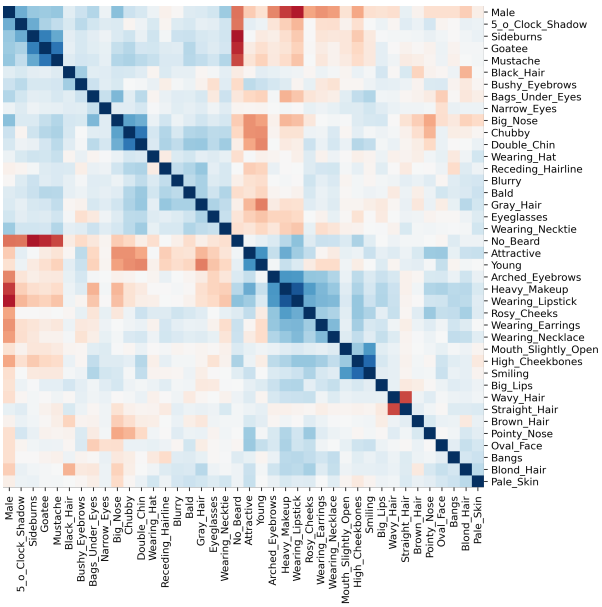


Figure 4: Latent semantic correlations of Diffusion Autoencoder. Blue indicates positive correlations and red indicates negative correlations.

One main challenge of manipulating multiple attributes is maintaining the unrelated attributes. For compositional generation, these are the out-of-scope attributes; for sequential editing, these also include the previously manipulated attributes. To tackle the challenge, previous methods either design a protection with linear manipulation such as InterfaceGAN (Shen et al., 2020), or use non-linear manipulation such as StyleFlow and LACE. As we have shown in previous sections, our linear method can be very competitive against other non-linear methods despite its simplicity.

To understand its power, we examine the linear classifiers that are learned for manipulation. Here we use the heatmap to visualize the correlation between each pair of linear classifiers learned on the semantic latent space of Diffusion Autoencoder, as shown in Figure 4. As we can see, the semantic latent space of Diffusion Au-

toencoder is favorably disentangled and thus the linear classifiers show strong orthogonality frequently. This means that even without a specific protection mechanism as in (Shen et al., 2020), our linear method is still capable of preserving the identity in most cases. In Table 3, we list the changes of conditional accuracy of other attributes when a single attribute is linearly manipulated. The small changes indicate that the attributes are well disentangled in such generative models, and further explain the efficacy of our linear method.

Table 3: Changes of conditional accuracy for each linear sequential edit

	yaw	smile	age	glasses
Edit-1	—	-0.001	+0.001	-0.002
Edit-2	+0.003	—	+0.001	+0.007
Edit-3	+0.001	+0.000	—	-0.013
Edit-4	+0.000	-0.012	-0.005	—

5.2 When is the non-linear method preferred?

Our linear method performing well does not diminish the value of non-linear manipulation methods. Abdal et al. (2021) argue that linear manipulation often moves a latent code outside the latent distribution which leads to low-quality generation, which is often the main motivation for non-linear methods. We agree that linear manipulation can generate sub-optimal results, but we argue that such a scenario could only happen when the original input already lies in a low density region. For most inputs, linear manipulation would satisfy as we have demonstrated; for novel combinations of attributes that are out-of-domain, non-linear manipulation might be necessary for obtaining reasonable, high-quality editions.

6 Conclusions

In conclusion, we propose a principled method for compositional generation and manipulation in the vision domain. Our method, built on latent diffusion models, is agnostic to different generative models and latent spaces. We show that our method performs competitively on various compositionality tasks, and is especially advantageous in real image manipulation. Future work will focus on modeling more complicated relations between attributes and achieving compositionality on more challenging datasets.

References

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: How to embed images into the StyleGAN latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4432–4441, 2019.

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8296–8305, 2020.

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.

Korbinian Abstreiter, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. Diffusion-based representation learning. *arXiv preprint arXiv:2105.14257*, 2021.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8188–8197, 2020.

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

Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation and inference with energy based models. *arXiv preprint arXiv:2004.06030*, 2020.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANspace: Discovering interpretable GAN controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

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

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

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. *arXiv preprint arXiv:2206.01714*, 2022.

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.

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

Weili Nie, Arash Vahdat, and Anima Anandkumar. Controllable and compositional generation with latent-space energy-based models. *Advances in Neural Information Processing Systems*, 34:13497–13510, 2021.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022.

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.

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

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a StyleGAN encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2287–2296, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

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.

Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1532–1540, 2021.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, 2020.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

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

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

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.

Zongze Wu, Dani Lischinski, and Eli Shechtman. StyleSpace analysis: Disentangled controls for StyleGAN image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12863–12872, 2021.

A APPENDIX

A.1 PROOF OF LEMMA 2

Proof. Lemma .2 can be proved using $p(z_{t-1}|z_t, y) = Zp(z_{t-1}|z_t)p(y|z_{t-1})$ (Z is a normalizing constant) and following the same routine as the proof of Lemma .1.

$$\log p(z_0, y) = \log \int p(z_{0:T}|y)p(y)dz_{1:T} \tag{9}$$

$$\geq \mathbb{E}_{q(z_{1:T}|z_0)} \log \frac{p(z_{0:T}|y)p(y)}{q(z_{1:T}|z_0)} \tag{10}$$

$$= \mathbb{E}_{q(z_{1:T}|z_0)} \left[\log \frac{p(z_T)}{q(z_T|z_0)} + \sum_{t=2}^{T} \log \frac{p(z_{t-1}|z_t, y)}{q(z_{t-1}|z_t, z_0)} + \log p(z_0|z_1, y) \right] + C_1 \tag{11}$$

$$= \mathbb{E}_{q(z_{1:T}|z_0)} \left[\log \frac{p(z_T)}{q(z_T|z_0)} + \sum_{t=2}^{T} \log \frac{p(z_{t-1}|z_t)}{q(z_{t-1}|z_t, z_0)} + \log p(z_0|z_1) \right. \tag{12}$$

$$\left. + \sum_{t=1}^{T} \log p(y|z_{t-1}) \right] + C_2 \tag{13}$$

$$= \mathcal{L}_{uncond} + \mathbb{E}_{q(z_{1:T}|z_0)} \left[\sum_{t=1}^{T} \log p(y|z_{t-1}) \right] + C_2. \tag{14}$$

For clarity purposes, we only show the proof with single condition y, but derivations can be easily extended to multiple y for compositional generation, and the cases with $\hat{z}$ for manipulation.

□

A.2 NETWORK ARCHITECTURE

We train a deep MLP with skip connections for our latent diffusion models, the same as the latent DDIM used in Diffusion Autoencoder. The latent diffusion model is composed of 10 sequential 2-layer MLP blocks. The input and output dimensions of the MLP blocks are 512 (decided by the latent dimension of StyleGAN2 and Diffusion Autoencoder), and all hidden dimensions in the middle are 2048. After the output layer of each MLP block, the time embeddings are injected by group normalization, followed by skip connections from the inputs of the first MLP block. The activation function is SELU. The time-embedding network is a 2-layer MLP that transforms the 64 dimensional sinusoidals to 512 dimensional time embeddings. The hidden layer is 512 dimensional, and the activation function is SELU.

A.3 TRAINING STRATEGIES

To train our latent diffusion models, we follow the same strategies as in (Nichol & Dhariwal, 2021; Preechakul et al., 2022).

Attribute	Validation Accuracy	Test Accuracy
Smile	92.00	91.67
Gender	93.40	94.20
Glasses	92.60	91.30
Beard	93.40	91.60
Hair color	75.40	75.50
Yaw	98.07	98.13
Age	93.37	93.64

Table 4: Validation and test accuracy of linear latent classifiers of StyleGAN2.

We train linear classifiers on the $\mathcal{W}_s$ space of StyleGAN2 using the same data from StyleFlow and LACE. We use the Adam optimizer to train each linear model for 200 epochs with an initial learning

rate of 1e-4, which is decreased by a factor of 0.3 at the 160 and 180 epochs. Note that different from LACE, we train binary classifiers for multi-class attributes. The training results are shown in Table. 4. We use pre-trained latent classifiers for Diffusion Autoencoders.