Exploring Compositional Visual Generation with Latent Classifier Guidance

Changhao Shi[1*]   Haomiao Ni[2*]   Kai Li[4]   Shaobo Han[4]   Mingfu Liang[3*]   Martin Renqiang Min[4]

[1]University of California, San Diego, CA, USA
[2]The Pennsylvania State University, University Park, PA, USA
[3]Northwestern University, Evanston, IL, USA
[4]NEC Laboratories America, Princeton, NJ, USA

[1]cshi@ucsd.edu   [2]hfn5052@psu.edu   [3]mingfuliang20202@u.northwestern.edu
[4]{kaili, shaobo, renqiang}@nec-labs.com

Abstract

Diffusion probabilistic models have achieved enormous success in the field of image generation and manipulation. In this paper, we explore a novel paradigm of using the diffusion model and classifier guidance in the latent semantic space for compositional visual tasks. Specifically, we train latent diffusion models and auxiliary latent classifiers to facilitate non-linear navigation of latent representation generation for any pre-trained generative model with a semantic latent space. We demonstrate that such conditional generation achieved by latent classifier guidance provably maximizes a lower bound of the conditional log probability during training. To maintain the original semantics during manipulation, we introduce a new guidance term, which we show is crucial for achieving compositionality. With additional assumptions, we show that the non-linear manipulation reduces to a simple latent arithmetic approach. We show that this paradigm based on latent classifier guidance is agnostic to pre-trained generative models, and present competitive results for both image generation and sequential manipulation of real and synthetic images. Our findings suggest that latent classifier guidance is a promising approach that merits further exploration, even in the presence of other strong competing methods.

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) [19], normalizing flows [31], and generative adversarial networks (GANs) [2, 7, 8, 11, 50], to the very recent diffusion probabilistic models [4, 14, 26, 33, 39, 40] and score-based models [41, 42], generating high-quality, real-istic images has become easier, if not impossible before. Despite the previous significant progress, 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 class of methods tackles the problem directly in the image space, either relying on energy-based models (EBMs) or drawing inspiration from them [10, 23]. Such technique is referred as classifier-free guidance in the literature, in contrast to classifier guidance that relies on auxiliary image classifiers [26]. Unlike classifier guidance that is mainly used for controllable generation with a single condition, the classifier-free guidance is naturally composable and suitable for multiple conditions. However, such methods can not leverage the nice properties of the latent space such as disentanglement [6]. Also, training multiple image space sub-models, either EBMs or diffusion models, can be cumbersome, especially when the number of conditions grows.

The second class of methods focuses on the latent space of pre-trained generative models. These methods aim to find 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 [17,18], linear control is possible by carefully identifying and combining the latent directions of each attribute [12, 36, 37, 50]. 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

---

*Work done during the internship at NEC Laboratories America.

new attributes [3] or not agnostic to various models and latent spaces [27]. Using diffusion models to control latent space through classifier-free guidance has been widely recognized [33], but not within the context of compositionality. Using diffusion models and classifier guidance in latent space however, is still a missing piece.

In this paper, we aim to fill in this missing piece and answer the question: is the latent diffusion model with latent classifier guidance useful for compositional image generation and manipulation[1]? We demonstrate that classifier guidance can help diffusion probabilistic models to manipulate latent spaces in a non-linear way, and this process can be further simplified to a linear version that resembles vector arithmetic-based manipulation with additional assumptions. 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. To facilitate the manipulation of synthetic or real images, we introduce an additional guidance term by framing the problem as incorporating a source image condition into the compositional generation process. We demonstrate that employing latent classifier guidance with diffusion models maximizes a lower bound of the conditional log probability function, providing a provable approach for conditioning on multiple attributes. Our experiments validate the effectiveness of this technique, as it can generate realistic images with various attribute compositions and manipulate both **synthetic and real** images in a coherent manner. We also find that the linear version of our proposed method based on vector arithmetic can serve as a strong baseline in many scenarios, despite previous studies focusing on non-linear manipulation.

2. Related Work

(Conditional) diffusion models. Diffusion models have become increasingly favorable over other generative models such as GANs [11] and VAEs [19] due to their photo-realistic generation quality and ease of training. [39] proposed the first functional framework of diffusion models from the perspective of thermodynamics, then this framework was followed by [14, 40–42] whose works established the foundation of diffusion models that we see today. For the conditional generation with diffusion models, [9] further formulated classifier guidance and lifted the generation quality of diffusion models over previous state-of-the-art GANs. [15] then proposed classifier-free guidance which nowadays is used in many large-scale image generation engine [25, 30, 35]. Although most of these diffusion models work on the image space, recently latent diffusion models

---

[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.

have also drawn great attention and achieved remarkable results [4, 33, 44].

Compositionality in latent space. Due to the wide success of StyleGANs [17, 18] 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. [36] found a direction for each attribute by training linear SVMs, then perturbed the latent point along the orthogonal projection of these directions to prevent unwanted semantic changes. [50] detected 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 [12], or the SVD of the first subsequent linear layer [37]. On the non-linear side, [3] designed a framework to control a set of pre-decided attributes using conditional normalizing flow. [27] used 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 tackled compositional generation in the image space. These methods either directly used EBMs or employed diffusion models that can be considered as EBMs. [10] trained EBMs for each condition and composed them together by defining new energy functions based on each individual energy function and their relations. [23] 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 [25, 30, 35] 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 [23].

3. Methodology

In this section, we will describe how the latent diffusion model and classifier guidance can be used to generate and manipulate images in a principled way.

3.1. Latent Diffusion Modeling

A diffusion model is a type of deep latent variable model that approximates an unknown data distribution $p(x)$ through smooth, iterative denoising steps. It maps a pre-defined noise distribution to the data distribution using the following formula: $p_\theta(x_0) = \int p_\theta(x_{0:T}) dx_{1:T}$, where $x_{1:T}$ are the latent variables with 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 $\mathcal{N}(x_T; \mathbf{0}, \mathbf{I})$. The reverse sampling 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 reconstruct the original images from the corrupted inputs. Specifically, for de-noising diffusion probabilistic models (DDPM) [14], the training objective is formulated as a re-weighted variational bound by treating DDPMs as VAEs, while for scored-based generative models [41], the objective is derived using score matching. Once trained, to generate samples from the learned distribution, one first samples x_T from a standard Gaussian and then uses the reverse process to transform it into the image space.

Here, we focus on leveraging the latent space of a pre-trained generative model. Specifically, we train a diffusion model to approximate the latent distribution $p(z)$ of a pre-trained generator G that maps a latent space $\mathcal{Z}$ to the image space $\mathcal{X}$. Modeling the latent space has several advantages over modeling the image space. For instance, the latent space enjoys properties such as disentanglement [6], which can facilitate more controllable manipulations of the generated images. Additionally, using various guidance techniques in the latent space is often more feasible since training latent guidance terms is generally easier than training other manipulation methods in image space [36].

3.2. Conditional and Compositional Generation

Conditional generation with diffusion models relies on perturbing unconditional generation with user-specified guidance terms, namely classifier guidance [9, 39, 42] and classifier-free guidance [15]. Although classifier-free guidance performs competitively in image space and is sometimes more favorable than classifier guidance [15, 23], we argue that using classifier guidance in latent diffusion models has its unique advantages. Regarding the under-performance of classifier guidance in the image space, one popular suspicion is that image classifiers tend to learn shortcuts from suspicious correlations. For example, a deep neural network classifier on the attribute "old" can be misguided by "white hair" and ignore its holistic features. This problem is alleviated in a compact, even disentangled latent space, if the semantic directions of 'old' and 'white hair' are orthogonal. Also, deep image classifiers are typically vulnerable to adversarial attacks, while latent classifiers with much few parameters suffer less from this problem. Another benefit is that classifiers are usually easier to train than

diffusion models used in classifier-free guidance. Finally, when the classifiers are linear, classifier guidance resembles linear arithmetic methods, as we will show in Section 3.4.

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

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

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

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

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 [39], 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 Eq. (1), the overall score function for manipulation then becomes

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

similarly for Eq. (2) with the presence of negation. When $p(\hat{z}|z_t)$ is modeled by an isotropic Gaussian distribution,

the new guidance term $\gamma_t \nabla_z \log p(\hat{z}|z)$ behaves as a regularization term $\nabla_{z_t} \|z_t - \hat{z}\|_2^2$.

3.3. Model Training

A true compositional model should be able to easily encompass new attributes without re-training the whole model. Indeed, 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 encompassing new attributes simply requires training classifiers on them, 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:*

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

See [14] for the detailed proof.

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

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

and with independent conditions $\{y^1, y^2, ..., y^n\}$ and $\hat{z}$, the ELBO is given by:

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

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.

$$
\log p(z_0, y)
$$
$$
= \log \int p(z_{0:T}|y)p(y)dz_{1:T}
$$
$$
\geq \mathbb{E}_{q(z_{1:T}|z_0)}\log \frac{p(z_{0:T}|y)p(y)}{q(z_{1:T}|z_0)}
$$
$$
= \mathbb{E}_{q(z_{1:T}|z_0)}\Bigg[\log \frac{p(z_T)}{q(z_T|z_0)} + \sum_{t=2}^{T}\log \frac{p(z_{t-1}|z_t, y)}{q(z_{t-1}|z_t, z_0)}
$$

$$
+ \log p(z_0|z_1, y)\Bigg] + C_1
$$
$$
= \mathbb{E}_{q(z_{1:T}|z_0)}\Bigg[\log \frac{p(z_T)}{q(z_T|z_0)} + \sum_{t=2}^{T}\log \frac{p(z_{t-1}|z_t)}{q(z_{t-1}|z_t, z_0)}
$$
$$
+ \log p(z_0|z_1) + \sum_{t=1}^{T}\log p(y|z_{t-1})\Bigg] + C_2
$$
$$
= \mathcal{L}_{uncond} + \mathbb{E}_{q(z_{1:T}|z_0)}\Bigg[\sum_{t=1}^{T}\log p(y|z_{t-1})\Bigg] + C_2.
$$

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. $\square$

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.

3.4. Connection to Linear Arithmetic

The regularized guidance 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 Eq. (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*

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

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 Eq. (3) should converge to 0 as follows:

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

which leads to the above analytic solution. $\square$

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 latent classifier guidance in later comparisons.

---

[2]Let the scalar absorbed by α_t^i.

4. Experiments

We evaluate classifier-guided latent diffusion models for compositional generation and manipulation tasks on two pre-trained models, StyleGAN2 [18] and Diffusion Autoencoder [29]. To use the described framework in the intermediate latent space ($\mathcal{W}_s$ space) of a pre-trained StyleGAN2, we first train latent DDIM [40] on 100,000 w_s vectors sampled from the push-forward distribution given by the style generation. We then train linear classifiers on the $\mathcal{W}_s$ space using the latent-label pairs provided by [3]. Note that although Eq. (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 in our preliminary experiments. The latent diffusion model is the same as the latent DDIM used in [29]. The performance of the classifiers can be found in Table 1. 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 [18] to get the initial latent space $\mathcal{Z}_s$ and intermediate $\mathcal{W}_s$ space encodings and then employ the pre-trained pSp encoder [32] to get $\mathcal{W}_s+$ space encodings, where $\mathcal{W}_s+$ is a concatenation of 18 different 512-dimensional w_s vectors in StyleGAN2. With Diffusion Autoencoder, we can directly use their pre-trained encoders to get semantic vectors.

Following [27], we consider three metrics for our evaluation: Fréchet Inception Distance (FID) [13], face identity loss (ID) [3] 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. A pair of input and manipulated face images are embedded by a pre-trained face recognition model[3], and the ID score is computed as the distance between their embeddings. 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 latent classifier guidance to generate images with multiple desired attributes. For high-resolution images (1024×1024), we select Style-GAN2 as the pre-trained generator; for low-resolution images (256×256), we use Diffusion Autoencoder.

We compare our proposed method, which we refer as LCG (Latent Classifier Guidance) from below, with Style-Flow [3] and LACE [27]. Results of StyleFlow are directly taken from [27] for the conjunction of "gender", "smile"

Table 1. Validation and test accuracy of linear latent classifiers of StyleGAN2.

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

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 as the $\mathcal{Z}_s$ or $\mathcal{W}_s$ space of StyleGAN2. However, we can still apply latent classifier guidance because it can be easily fitted by diffusion models.

The quantitative comparison is shown in Table 2 and the qualitative comparison shown in Fig. 1 and Fig. 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, with latent classifier guidance, using simple linear arithmetic (LCG-Linear) and latent diffusion models (LCG-Diffusion) both perform competitively against previous non-linear methods.

4.2. Compositional Manipulation

We evaluate latent classifier guidance 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 [27] 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 3. Note that both of the 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. Latent classifier guidance 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 [1] and $\mathcal{S}$ space [50] are better choices for real image editing. However, LACE is restricted to the

(a) LACE-LD [27]

(b) LACE-ODE [27]

(c) LCG-Linear (Ours)

(d) LCG-Diffusion (Ours)

Figure 1. Qualitative comparison among different methods for compositional generation based on the latenet space of the pre-trained StyleGAN2 generator under the resolution of 1024×1024. The target attribute conditions are: female, smile, and 55 y/o (i.e., "years old").
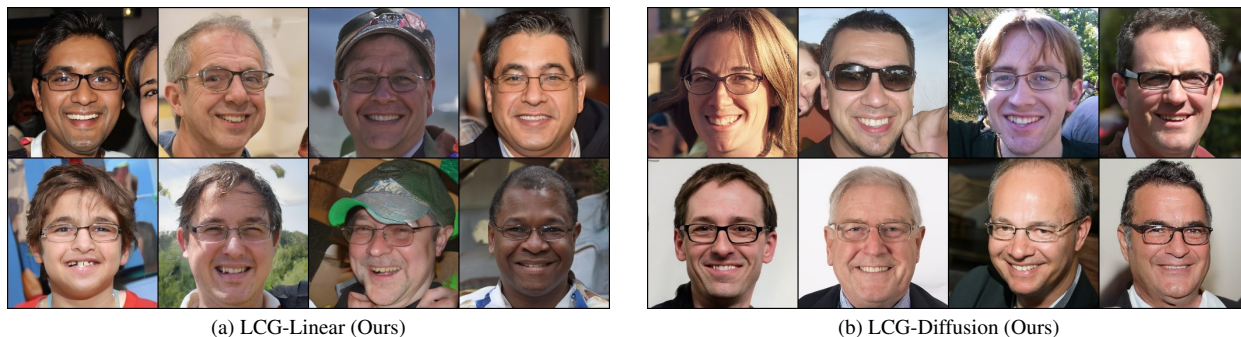

(a) LCG-Linear (Ours)

(b) LCG-Diffusion (Ours)

Figure 2. Qualitative comparison of our methods for compositional generation based on the latent space of Diffusion Autoencoder generator under the resolution of 256×256. The target attribute conditions are: male, smile, and glasses.

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. Latent diffusion models, on the other hand, can be trained on either existing or new expanded spaces, where the semantics are richer and the inversion is easier.

As shown in Figure 3, latent classifier guidance 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, the latent classifier guidance only requires inversion into $\mathcal{W}_s$ or $\mathcal{W}_s+$ space and controls attributes better as well as preserves the identity more faith-

fully 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 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?

One main challenge of manipulating multiple attributes is maintaining the non-targeted attributes. For composi-

Table 2. Quantative comparison of different methods for compositional generation based on the latent space of pre-trained StyleGAN2.

| Method | gender, smile, age | | | | -gender, smile, -haircolor | | | |
| | FID ↓ | ACC ↑ | | | FID ↓ | ACC ↑ | | |
		gender	smile	age		gender	smile	haircolor
StyleFlow [3]	43.88	0.718	0.870	0.874	—	—	—	—
LACE-LD [27]	22.34	0.953	0.954	**0.925**	**22.86**	0.678	0.958	0.924
LACE-ODE [27]	**22.03**	0.964	0.967	**0.925**	23.51	0.649	0.970	0.935
LCG-Linear (Ours)	22.46	0.980	**0.982**	0.863	23.94	0.948	**0.995**	**0.936**
LCG-Diffusion (Ours)	26.49	**0.981**	0.968	0.863	29.62	**0.987**	0.954	0.906

Table 3. Quantative comparison of different methods for sequential editing with StyleGAN2.

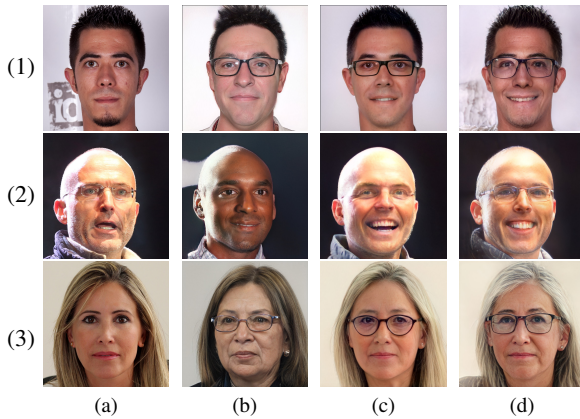| Method | FID ↓ | ID ↓ | ACC ↑ | | | |
			yaw	smile	age	glasses
StyleFlow [3]	44.13	0.549	**0.947**	0.773	0.817	0.876
LACE-ODE [27]	27.49	0.501	0.938	0.956	**0.881**	**0.997**
LCG-Linear (Ours)	29.48	**0.290**	0.887	**0.983**	0.875	0.786
LCG-Diffusion (Ours)	**24.06**	0.445	0.903	0.963	0.845	0.843

Figure 3. Qualitative comparison among different compositional manipulation methods on real image inputs. Target attributes from the top row to the bottom row are (1) glasses, smile, 55 y/o, (2) no glasses, smile, 28 y/o, (3) glasses, white hair, right face. Each column from left to right are (a) original images, (b) LACE-ODE [27], (c) linear latent classifier guidance in $\mathcal{W}_s$, (d) linear latent classifier guidance in $\mathcal{W}_s$+.
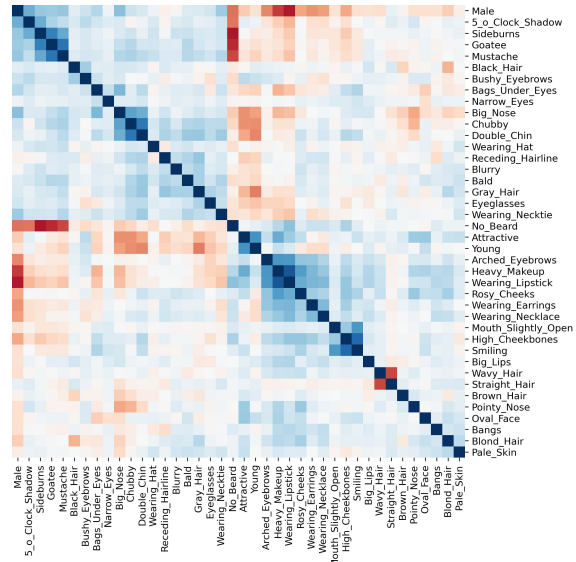

Figure 4. Visualization of Latent semantic correlations of Diffusion Autoencoder. Bluer regions indicate positive correlations and redder areas suggest negative correlations.

tional 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 [36], or use non-linear manipulation such as StyleFlow and LACE. As we have shown in previous sections, our LCG-linear 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 Fig. 4. As we can see, the semantic latent space of Diffusion Autoencoder is favorably disentangled and thus the linear classifiers show strong orthogonality frequently. This means that even without a specific protection mechanism as in [36], LCG-linear is still capable of preserving the identity in most cases. In Table 4, we list the changes of conditional accuracy of other attributes when a single attribute is linearly manipulated. The

four edits correspond to the four attributes "yaw", "smile", "age", and "glasses" respectively. The small changes indicate that the attributes are well disentangled in such generative models, and further explain the efficacy of LCG-linear.

Table 4. Changes of conditional accuracy for each linear sequential editing.

	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?

Despite the complicity of the non-linear diffusion based method, it does not always perform favorably against the linear version. The main motivation for non-linear methods, as argued in [3], is that linear manipulation often moves a latent code outside the latent distribution which leads to low-quality generation. Indeed, in linear manipulation, we assume a non-informative $p(z_t)$ which favors different regions of sample space equally, regardless of the actual latent distribution. This indicates that non-linear control is likely to profit when the generation needs to traverse low density region or is simply out-of-distribution.

An example is sequential editing. Sequential editing is more prone to low density region, as the new edit is conditioned on previous edits that possibly have already guided the latent to low density regions. To see this, imagine a 3-D Guassian distribution where each axis represents an attribute and we want to guide a sample point from $[-1, -1, -1]$ to $[1, 1, 1]$. Compositional generation is analogous to a direct path $[-1, -1, -1] \rightarrow [1, 1, 1]$, while sequential editing is analogous to a path $[-1, -1, -1] \rightarrow [-1, -1, 1] \rightarrow \rightarrow [-1, 1, 1] \rightarrow [1, 1, 1]$ that traverses more low density regions. As we can in Table 3, LCG-diffusion outperforms LCG-linear on FID, generating more realistic images. Characterizing the latent distribution with diffusion model in this case is favorable than a non-informative one, as the diffusion model always pulls the sample toward high density region thus preventing it going out-of-distribution. A downside of this is that, as we can see from the ID score, keeping images realistic is at the cost of losing identity preservation. Improving the identity preservation should serve as an interesting topic for future exploration.

6. Conclusion and Future Work

In conclusion, we study the efficacy of using latent diffusion models and latent classifier guidance for compositional visual generation and manipulation. Specifically, we train latent diffusion models and auxiliary latent classifiers

to facilitate non-linear navigation of latent representation generation for two pre-trained generative models, Style-GAN2 and Diffusion Autoencoder. We demonstrate such a paradigm is suitable for compositional visual tasks both theoretically and empirically. Our findings suggest that latent classifier guidance is a promising approach that deserves further research, even in the presence of other strong methods such as Stable Diffusion [33].

In our future work, we plan to explore modeling more complicated relations between attributes and aim to achieve compositionality on more challenging datasets and tasks, such as text/class-conditioned video generation [5, 16, 24, 38]. We are also interested in the performance of latent classifier guidance in out-of-distribution settings. In addition, we acknowledge that relying on a pretrained generative model with a semantic latent space may not be practical or feasible in some scenarios. To address this potential problem, we will explore the possibility of reorganizing the latent space of the pretrained generative model to construct a semantic latent space in the post-pretraining stage. Also, the demand for generating unseen classes and unseen subconcept of an existing class is of surged interest in the community [20, 21, 28, 34]. To tackle these new challenges for compositional generation, we plan to investigate the applicability of leveraging the continual learning and incremental learning techniques [22, 43, 45–49] to extend the semantic latent space of the generative model accordingly.

References

[1] 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*, pages 4432–4441, 2019. 5

[2] 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*, pages 8296–8305, 2020. 1

[3] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021. 2, 5, 7, 8

[4] Korbinian Abstreiter, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. Diffusion-based representation learning. *arXiv preprint arXiv:2105.14257*, 2021. 1, 2

[5] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, volume 1, page 2, 2019. 8

[6] 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. 1, 3

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

In *International Conference on Learning Representations*, 2018. 1

[8] 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*, pages 8188–8197, 2020. 1

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2, 3

[10] Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation and inference with energy based models. *arXiv preprint arXiv:2004.06030*, 2020. 1, 2

[11] 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. 1, 2

[12] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANspace: Discovering interpretable GAN controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020. 1, 2

[13] 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. 5

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2, 3, 4

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3

[16] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 8

[17] 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*, pages 4401–4410, 2019. 1, 2

[18] 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*, pages 8110–8119, 2020. 1, 2, 5

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2

[20] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. *arXiv preprint arXiv:2303.13516*, 2023. 8

[21] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022. 8

[22] Mingfu Liang, Jiahuan Zhou, Wei Wei, and Ying Wu. Balancing between forgetting and acquisition in incremental subpopulation learning. In *Computer Vision–ECCV 2022:*

17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 364–380. Springer, 2022. 8

[23] 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. 1, 2, 3

[24] Haomiao Ni, Changhao Shi, Kai Li, Sharon X. Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models, 2023. 8

[25] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[26] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1

[27] 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. 2, 5, 6, 7

[28] Yotam Nitzan, Michaël Gharbi, Richard Zhang, Taesung Park, Jun-Yan Zhu, Daniel Cohen-Or, and Eli Shechtman. Domain expansion of image generators. *arXiv preprint arXiv:2301.05225*, 2023. 8

[29] 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*, pages 10619–10629, 2022. 5

[30] 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. 2

[31] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 1

[32] 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*, pages 2287–2296, 2021. 5

[33] 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*, pages 10684–10695, 2022. 1, 2, 8

[34] 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. 8

[35] 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. 2

[36] 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*, pages 9243–9252, 2020. 1, 2, 3, 7

[37] 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*, pages 1532–1540, 2021. 1, 2

[38] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 8

[39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1, 2, 3

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2, 5

[41] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 3

[42] 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*, 2020. 1, 2, 3

[43] Shengbang Tong, Xili Dai, Ziyang Wu, Mingyang Li, Brent Yi, and Yi Ma. Incremental learning of structured memory via closed-loop transcription. In *The Eleventh International Conference on Learning Representations*, 2023. 8

[44] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021. 2

[45] Gido M Van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1):4069, 2020. 8

[46] Riccardo Volpi, Diane Larlus, and Grégory Rogez. Continual adaptation of visual representations via domain randomization and meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4443–4453, 2021. 8

[47] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 8

[48] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 631–648. Springer, 2022. 8

[49] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022. 8

[50] 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*, pages 12863–12872, 2021. 1, 2, 5