Unsupervised Controllable Generation with Score-based Diffusion Models: Disentangled Latent Code Guidance

Yeongmin Kim
KAIST
alsdudrla10@kaist.ac.kr

Dongjun Kim
KAIST
dongjoun57@kaist.ac.kr

HyeonMin Lee
KAIST
leehm1111@kaist.ac.kr

Il-chul Moon
KAIST, Summary.AI
icmoon@kaist.ac.kr

Abstract

From the impressive empirical success of Score-based diffusion models, it is recently spotlighted in generative models. In real-world applications, the controllable generation enriches the impact of diffusion models. This paper aims to solve the challenge by presenting the method of control in an unsupervised manner. We propose the Latent Code Guidance Diffusion Model (LCG-DM), which is the first approach to apply disentanglement on Score-based diffusion models. Disentangled latent code can be considered as a pseudo-label, since it separately expresses semantic information in each dimension. LCG-DM is a Score-based diffusion model that reflects disentangled latent code as the condition. LCG-DM shows the best performance among baselines in terms of both sample quality and disentanglement on dSprites dataset. LCG-DM can manipulate images on CelebA dataset, with comparable FID performance compared to non-disentangling Score-based diffusion models. Furthermore, we provide experimental results of scaling method that reflects more on pseudo-label with MNIST dataset.

1 Introduction

Deep generative models (DGMs) estimate the underlying data distribution from a finite number of observed data samples [23]. This estimation comes from the iterative minimization of the statistical divergence between the model distribution and the data distribution [16]. One branch of DGM starts the divergence minimization by explicitly specifying the model distribution and its transformations, i.e. autoregressive models [18], Variational Autoencoder (VAE) [12, 4, 28], Flow-models [21], Score-based diffusion models [27, 24, 10], etc, and this paper refers to these DGM variants as likelihood-based models. Among them, the Score-based diffusion models have been recently developed with the high quality of sampled data instances, which Generative Adversarial Network(GAN), a likelihood-free model, had been considered the best approach. Given the better sample quality with an explicit density model by the Score-based diffusion models, researchers are expanding research questions on diffusion models to control [19, 22], guide [5, 8], and disentangle the model distribution while maintaining the sample quality.

The control and the disentanglement on the model distribution are enabling factors to utilize DGM in many industrial applications [20, 6]. Meanwhile, the controllable generation [1, 22] requires complete joint annotations on data instances, so the joint condition will specify the correlation between individual condition random variables. Because of the large joint condition space and limited

Score-based Method Workshop at Conference on Neural Information Processing Systems (NeurIPS 2022).

data samples to sufficiently shape the density on the condition space, the disentangled representation learning (DRL) on the condition variables makes controllable generation feasible with limited data instances [7, 11, 2, 3, 13]. Moreover, DRL enables synthesizing a data instance on a certain condition dimension with minimal impact on the other conditions. This disentanglement statistically results in independence across the condition random variables, and the learning objectives of DRL require mechanisms to ensure such independence.

There are several approaches of DRL, mostly originating from VAE [7, 11, 2] and GAN [3, 13] frameworks. In VAE, latent variables are regularized to be dimension-wise independent when they are inferred. However, the regularization for the disentanglement significantly reduces sample quality. In GAN, the regularization by mutual information makes the latent code disentangled, yet this regularization on GAN still degrades the sample quality with further learning instability.

Given the advanced sample quality from the Score-based diffusion model, it is natural progress to formulate a DRL variation of diffusion models. This paper proposes Latent Code Guidance Diffusion Model (LCG-DM), enabling unsupervised controllable sample generation, which is the first DRL model in the community of Score-based diffusion models.

2 Preliminary

2.1 Score-based diffusion models

Score-based diffusion models consist of forward and reverse diffusion processes. The forward diffusion process refers to adding noise to data, which is defined as Eq. (1) with a drift function, $\mathbf{f}(\mathbf{x}_t, t)$ and a volatility function, $g(t)$. $\mathbf{x}_0$ denotes a data random variable and $\mathbf{x}_t$ denotes a perturbed random variable at t. $\mathbf{w}_t$ refers to the standard Wiener process.

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t \tag{1}$$

While the forward process is a simple noise additive process, diffusion models define the reverse diffusion to denoise an image from a noisy input. Given that reverse diffusion is the reverse process of the forward diffusion, Eq. (2) derives the reverse diffusion. $p_t(\mathbf{x}_t)$ is a probability distribution of $\mathbf{x}_t$.

$$\mathrm{d}\mathbf{x}_t = [\mathbf{f}(\mathbf{x}_t, t) - g^2(t)\nabla \log p_t(\mathbf{x}_t)]\mathrm{d}\bar{t} + g(t)\mathrm{d}\overline{\mathbf{w}}_t \tag{2}$$

The score function $\nabla \log p_t(\mathbf{x}_t)$ has every information to reconstruct the data distribution after the reverse process. The estimation of the score function $\mathbf{s}_\theta(\mathbf{x}_t, t) \approx \nabla \log p_t(\mathbf{x}_t)$ is the goal of training.

$$\mathcal{L}_{dsm}(\boldsymbol{\theta}; \lambda) = \frac{1}{2}\mathbb{E}_t\big[\lambda(t)\mathbb{E}_{p(\mathbf{x}_0), p_{0t}(\mathbf{x}_t|\mathbf{x}_0)}[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \nabla \text{log} p_{0t}(\mathbf{x}_t|\mathbf{x}_0)\|_2^2]\big] \tag{3}$$

Since it is intractable to obtain the score $\nabla \log p_t(\mathbf{x}_t)$ from $\mathbf{x}_t$, some approaches approximate the score function with $\mathcal{L}_{dsm}$ [25, 26], where $\lambda(t)$ is a weighting function.

2.2 Disentangling Variational Autoencoders

A typical approach to disentangle latent variables in VAE is regularizing variational distribution $q_\phi(\mathbf{z})$, so $\mathbf{z}$ could be dimension-wise independent. The simplest approach is β-VAE to place a higher coefficient on $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}_0)||p(\mathbf{z}))$ in the evidence lower-bound (ELBO), so the Kullback–Leibler divergence enforces $q_\phi(\mathbf{z}|\mathbf{x}_0)$ to be closer to $N(\mathbf{0}, \mathbf{I})$, which is given as the prior distribution of $p(\mathbf{z})$. Later, Factor-VAE [11] and β-TCVAE [2] proposed Eq. (4) with $\gamma > 1$ showing better trade-off curve of the reconstruction sample quality and the disentanglement strength, where $\overline{q}_\phi(\mathbf{z}) = \prod_{j=1}^{d} q_\phi(z_j)$.

$$\mathcal{L}_{fvae}(\phi, \psi) = -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_0)}[\text{log} p_\psi(\mathbf{x}_0|\mathbf{z})] + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}_0)||p(\mathbf{z})) + \gamma D_{KL}(q_\phi(\mathbf{z})||\overline{q}_\phi(\mathbf{z})) \tag{4}$$

While we inherit the idea of disentangling regularization from Eq. (4) with hyper-parameter γ, we do not impose them directly on the loss function of Eq. (4) for a generation. Instead, LCG-DM creates a structure that uses inferred $\mathbf{z}$ as a condition in Score-based diffusion models.

3 Method

Figure 1 shows the overall structure of LCG-DM. $q_\phi(\mathbf{z}|\mathbf{x}_0)$ maps from $\mathbf{x}_0$ to disentangled latent code $\mathbf{z}$. The disentangled latent code vector $\mathbf{z}$ expresses semantic information with the independent

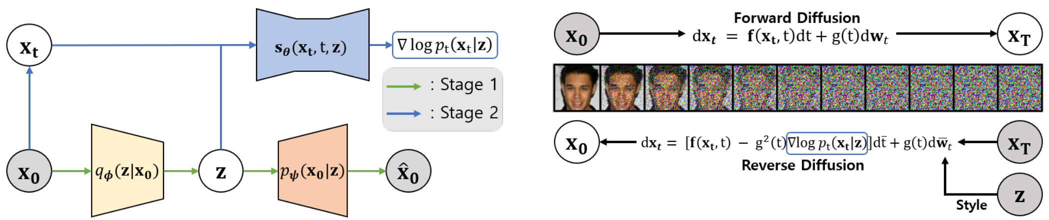

$\mathbf{x_t}$
$\mathbf{s}_\theta(\mathbf{x_t}, t, \mathbf{z})$
$\nabla \log p_t(\mathbf{x_t}|\mathbf{z})$
: Stage 1
: Stage 2
$\mathbf{x_0}$
$q_\phi(\mathbf{z}|\mathbf{x_0})$
$\mathbf{z}$
$p_\psi(\mathbf{x_0}|\mathbf{z})$
$\hat{\mathbf{x}}_0$
Forward Diffusion
$\mathbf{x_0}$
$d\mathbf{x}_t = \mathbf{f}(\mathbf{x_t}, t)dt + g(t)d\mathbf{w}_t$
$\mathbf{x_T}$
$\mathbf{x_0}$
$d\mathbf{x}_t = [\mathbf{f}(\mathbf{x_t}, t) - g^2(t)\boxed{\nabla \log p_t(\mathbf{x_t}|\mathbf{z})}]d\bar{t} + g(t)d\bar{\mathbf{w}}_t$
$\mathbf{x_T}$
Reverse Diffusion
Style
$\mathbf{z}$

Figure 1: Overall structure of proposed model: Latent Code Guidance Diffusion Model (LCG-DM). The left figure indicates the feed-forward structure and the right figure indicates the diffusion process.

dimensions of $\mathbf{z}$. From an alternative perspective, $\mathbf{z}$ becomes a pseudo-label of $\mathbf{x}_0$ in the ideal setting. Hence, we design LCG-DM to be a $\mathbf{z}$-conditional Score-based diffusion model with the disentangling mechanism.

In Figure 1, $\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z})$ is the function to estimate the score given $\mathbf{z}$, so $\mathbf{s}_\theta$ becomes the score estimation for learning. Meanwhile, the disentangled latent code of $\mathbf{z}$ is not provided from the dataset, so we create a VAE structure to extract the disentangled $\mathbf{z}$ from q_ϕ and p_ψ. This VAE structure follows Factor-VAE, and we utilize Eq. (4) in LCG-DM, which means that we have γ-controlled disentanglement term.

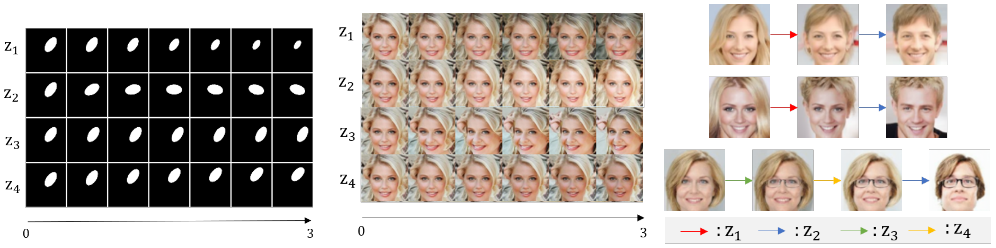
Since we leverage the disentangled latent condition $\mathbf{z}$ in training the diffusion model of $\mathbf{s}_\theta$, it is necessary to have stable signals from $\mathbf{z}$ in the diffusion model learning. Therefore, it is natural to have a step-wise approach in training.

$$\mathcal{L}_{overall}(\psi, \phi, \theta) = \mathcal{L}_{fvae}(\psi, \phi) + \mathcal{L}_{cdsm}(\phi, \theta) \tag{5}$$

In detail, $\mathcal{L}_{overall}$ in Eq. (5) denotes the loss function of LCG-DM. We conducted the training using two stages. In Stage 1, the model optimizes $\mathcal{L}_{fvae}$, so that q_ϕ extracts disentangled latent code $\mathbf{z}$. Once we have convergence in obtaining $\mathbf{z}$, we initiate Stage 2, which optimizes $\mathcal{L}_{csdm}$, so that $\mathbf{s}_\theta$ estimates the score function of conditional distribution $\nabla \text{log} p_t(\mathbf{x}_t|\mathbf{z})$. We derive that the minimization of conditional denoising score matching $\mathcal{L}_{cdsm}$ equals L2 minimization of conditional score matching in Appendix A.5.

$$\mathcal{L}_{cdsm}(\phi, \theta; \lambda) = \frac{1}{2}\mathbb{E}_t[\lambda(t)\mathbb{E}_{p(x_0), p_{0t}(x_t|x_0), q_\phi(z|x_0)}[\|\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) - \nabla \text{log} p_{0t}(\mathbf{x}_t|\mathbf{x}_0)\|_2^2]] \tag{6}$$

4 Results



(a) (b) (c)

Figure 2: Latent traversals on dSprites (a), and CelebA (b, c) datasets. The leftmost image in each row from $\mathbf{z} = \overrightarrow{0}$, and $\mathbf{x}_T$ is fixed in (a), (b). (a) shows each latent dimension corresponds to size, angle, X-axis, and Y-axis. (b) shows each latent dimension corresponds to age, hair brightness, azimuth, and face brightness. (c) shows different leftmost images since they sampled from different $\mathbf{x}_T$. Each colored arrow indicates changing of each latent dimension, and corresponding factors are changed. It shows that the image can be edited as desired.

We experimented with two benchmark datasets: dSprites [15] and CelebA [14]. dSprites is chosen because of its synthetic nature, so we can comprehend and measure the disentangled conditional latent variable $\mathbf{z}$, and its quality. CelebA provides real-world images for generation tasks by manipulating the disentangled latent variables.

4.1 dSprites dataset

dSprites is a synthetic dataset that has known six latent factors, i.e. shape, scale, position, etc, in generating a single image [15]. Therefore, we can explicitly measure the disentanglement quality by FVM [11] and MIG [2]. We compared the performance of LCG-DM against the disentangling structures of 1) VAE variations (β-VAE [7] and FactorVAE [11]), and 2) GAN variations (InfoGAN [3] and ID-GAN [13]). Additionally, we included DDPM++ +ST [10], which is the backbone diffusion model we use. DDPM++ +ST is not designed for disentanglement while the proposed model, LCG-DM, is the first disentanglement diffusion model to our knowledge.

Model	FID(↓)	FVM(↑)	MIG(↑)
VAE	74.10	0.58	0.05
β - VAE	130.01	0.60	0.23
β - TCVAE	143.99	0.67	0.23
FactorVAE	132.62	0.71	0.29
InfoGAN	8.74	0.47†	0.01†
ID-GAN	2.00†	0.65†	0.28†
DDPM++ +ST	1.28	N/A	N/A
LCG-DM	**1.01** (encoded)	**0.84**	**0.34**
	1.03 (prior)	N/A	N/A

Table 1: Comparison of FID and disentangling metrics on dSprites dataset. The symbol † denotes the results from other literature.

Table 1 shows the evaluation result from two perspectives: image generation and disentanglement. LCG-DM shows the best image generation in FID, and LCG-DM also provides the most disentangled latent variable $\mathbf{z}$ in FVM and MIG. Particularly, we experimented in two versions of LCG-DM: the prior version draws a sample from $\mathbf{z}, \mathbf{x}_T \sim N(\mathbf{0}, \mathbf{I})$, and the encoded version draws a sample from $\mathbf{z} \sim q_\phi$ and $\mathbf{x}_T \sim N(\mathbf{0}, \mathbf{I})$. The encoded version is slightly better because it takes advantage in drawing the condition with a sample image $\mathbf{x}_0$ via encoder q_ϕ. The outperformed FID is achieved by diffusion model structure, and the proposed conditioning structure makes higher γ minimally damages the generation.

γ	FID(↓)	FVM(↑)	MIG(↑)
0	0.75	0.40	0.01
10	1.38	0.58	0.26
20	1.03	0.84	0.34

Table 2: Sample quality and disentangling metrics by changing of γ on dSprites dataset.

Table 2 shows the impact of γ, which is a strength hyperparameter for disentanglement. As γ increases, FVM and MIG become better as expected. However, FID seems to be uncorrelated with γ, and the reason could be the separated training stages between the disentanglement by VAE and the generation by a diffusion model.

4.2 CelebA dataset

Table 3 shows the FID evaluation across disentangling models. We observe LCG-DM beats all baselines in its generation at the cost of encoded sampling. However, the gap between prior and q_ϕ yields a large discrepancy, and we attribute this to the inflexibility of the VAE structure. The large variational gap is observed not in dSprites dataset, but in CelebA datasets. This is because CelebA contains more complex information, so latent space also be more complex.

Model	FID(↓)
VAE	132.80
β - VAE	136.23
β - TCVAE	139.07
FactorVAE	134.52
GAN	3.34†
InfoGAN	4.93†
ID-GAN	4.08†
LCG-DM	**2.57** (encoded)
	10.65 (prior)

Table 3: FID comparison on CelebA (64× 64) dataset.

Table 4 shows the FID evaluation across diffusion models without disentanglement, except LCG-DM. While LCG-DM shows the worst FID from the prior version, it should be noted that 1) the NLL is comparable; and 2) LCG-DM has the disentangled latent variables unlike NCSN++ [27] and DDPM++ +ST [10].

Finally, Figure 2 shows the manipulation of a disentangled latent in image generations. Figure 2-(b) shows that the independent factor is smoothly changed for each $\mathbf{z}$ dimension. Figure 2-(c) shows that editing of image by adjusting each dimension. The same arrow means the change of the same dimension.

Model	FID(↓)	NLL(↓)
NCSN++	3.95†	2.39†
DDPM++(FID) +ST	**1.90**†	2.10†
DDPM++(NLL) +ST	2.90†	**1.96**†
LCG-DM	2.57(encoded)	2.29
	10.65(prior)	

Table 4: FID and NLL comparison with non-disentangling Diffusion Models on CelebA (64× 64) dataset.

5 Conclusion

This paper presents the first disentanglement diffusion model, LCG-DM. The proposed model utilizes a conditioning structure, so the impact of the disentanglement constraint minimally damages the image generation performance. This disentanglement diffusion model is a gateway to causal inference and counterfactual generations for various machine learning tasks, i.e. fairness research.

6 Acknowledgements

This research was supported by Research and Development on Key Supply Network Identification, Threat Analyses, Alternative Recommendation from Natural Language Data(NRF) funded by the Ministry of Education(2021R1A2C200981612)

References

[1] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021.

[2] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

[3] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

[4] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

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

[6] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022.

[7] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

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

[9] Hongxiang Jiang, Jihao Yin, Xiaoyan Luo, and Fuxiang Wang. Inference-infogan: Inference independence via embedding orthogonal basis expansion. *arXiv preprint arXiv:2110.00788*, 2021.

[10] Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. In *International Conference on Machine Learning*, pages 11201–11228. PMLR, 2022.

[11] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

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

[13] Wonkwang Lee, Donggyun Kim, Seunghoon Hong, and Honglak Lee. High-fidelity synthesis with disentangled representation. In *European Conference on Computer Vision*, pages 157–174. Springer, 2020.

[14] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[15] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[16] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016.

[17] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022.

[18] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pages 4055–4064. PMLR, 2018.

[19] 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.

[20] 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.

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

[22] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[23] Ruslan Salakhutdinov. Learning deep generative models. *Annual Review of Statistics and Its Application*, 2:361–385, 2015.

[24] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428, 2021.

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

[26] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pages 574–584. PMLR, 2020.

[27] 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.

[28] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020.

A Appendix

A.1 Scaling method

It is also an important problem to reflect the condition well in the generation, in conditional diffusion models. The estimation target of the conditional diffusion model is the score function of conditional probability $\nabla \text{log} p_t(\mathbf{x}_t|\mathbf{z})$. Since $\nabla \text{log} p_t(\mathbf{x}_t|\mathbf{z}) = \nabla \text{log} p_t(\mathbf{x}_t) + \nabla \text{log} p_t(\mathbf{z}|\mathbf{x_t})$, some paper propose separately estimating 1) an unconditional score function and 2) an additional network, which would become a classifier when $\mathbf{z}$ is regarded as a class label [27]. On top of this structure separation between $\nabla \text{log} p_t(\mathbf{x}_t)$ and $\nabla \text{log} p_t(\mathbf{z}|\mathbf{x}_t)$, recent works propose a scaling method that uses $\nabla \text{log} p_t(\mathbf{x}_t) + \lambda \cdot \nabla \text{log} p_t(\mathbf{z}|\mathbf{x}_t)$ as a conditional score function to strengthen the conditional input in the generation process [5].

Instead of separating and utilizing an adhoc network of $\nabla \log p_t(\mathbf{z}|\mathbf{x}_t)$, we adopt the classifier-free guidance method without the adhoc network. Subsequently, we estimate $\nabla \log p_t(\mathbf{x}_t) \approx \mathbf{s}_\theta(\mathbf{x}_t, t, \emptyset)$ and $\nabla \log p_t(\mathbf{x}_t|\mathbf{z}) \approx \mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z})$, which would turn our conditional score function into Eq. (7) by following the prior work of [8].

$$\nabla \log p_t(\mathbf{x}_t|\mathbf{z}) \approx (1+\lambda) \cdot \mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) - \lambda \cdot \mathbf{s}_\theta(\mathbf{x}_t, t, \emptyset) \tag{7}$$

Figure 3 shows the effects of scaling method. We observed that higher λ makes more reflection of pseudo-label, but has less diversity.

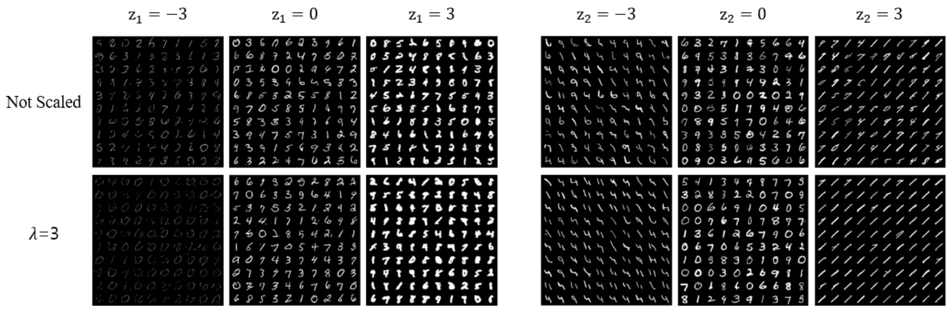


Figure 3: Generation results on MNIST dataset using scaling method. The first row is not scaled, and the second row is scaled using λ=3. Each cell shows random 100 samples, only fixing one dimension of $\mathbf{z}$. Scaled samples more reflect the condition z_1= slope, z_2 = thickness, but shows less diversity.

A.2 Additional latent traversals on CelebA dataset

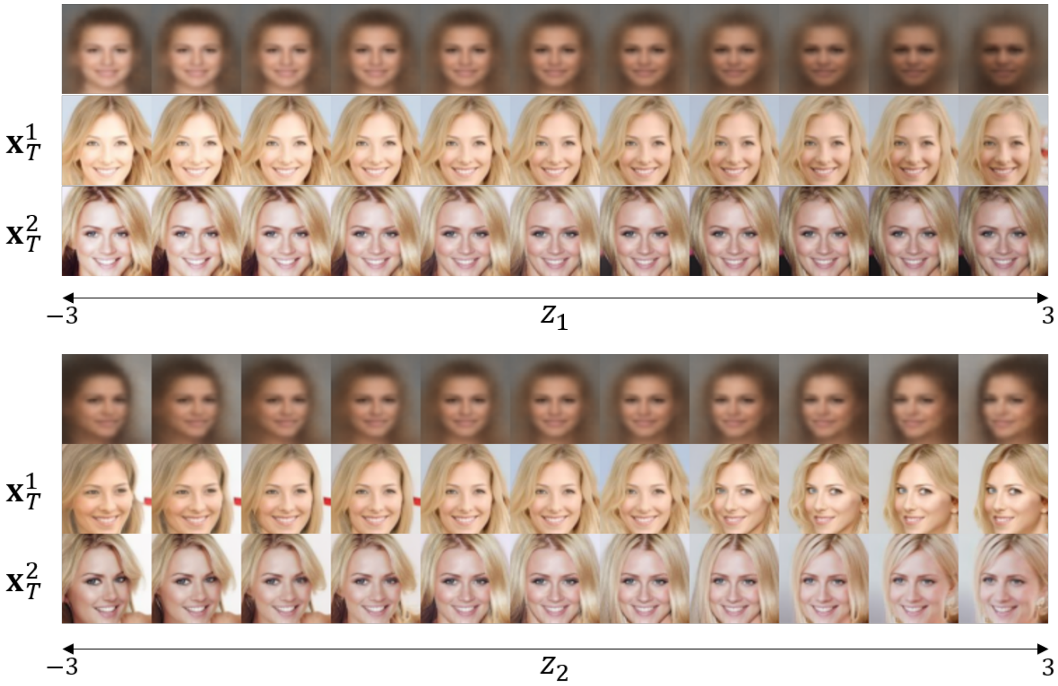


Figure 4: Sample comparison of Factor-VAE and LCG-DM which share the same latent code. LCG-DM shows superior sample quality and multi-modality from stochasticity of $\mathbf{x}_T$.

A.3 Experiment details

LCG-DM

The disentangling structure in LCG-DM uses Factor-VAE with γ=20, and 10 latent dimensions. The score network is pretrained in just the same setting as DDPM++ +ST [10], and fine-tuned with conditional embedding. We take 150,000 iterations with 64 batch sizes for fine-tuning. We use a probability flow ode sampler for a consistent sample from fixed $\mathbf{x}_T$ [27].

Disentangling VAES

We measured FID, FVM, and MIG by experimenting 3 times for VAE, β-VAE, Factor-VAE, and β-TCVAE. We use the same neural network architecture for encoder and decoder structures, which consists of stacked CNN and MLP layers. The latent dimension was set to 10, and the disentangling coefficient was set to 10.

Results from literature

We denote † on the table, which means the value from other literature. We refer to the results of GAN variants from [13, 9], and Score-based diffusion model variants from [10, 27].

A.4 Metrics

FID

Fréchet Inception Distance (FID) indicates the distance between training data and generated data, which is the most famous metric of sample quality. We use the clean-FID calculation module [17] with the Inception V3 network. We measure on 50k random samples in all settings.

NLL

We measure the Negative Log Likelihood (NLL) of LCG-DM on CelebA test dataset, and we compare it with non-disentangled diffusion models in Table 3. To measure it in LCG-DM, we calculate the upper bound of NLL as a state in Eq. (8). We compute $\log p_\theta(\mathbf{x}|\mathbf{z})$ using a probability flow ode [10].

$$\begin{aligned}
-\log p_\theta(\mathbf{x}) &= -\log \int_{\mathbf{z}} p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} \\
&= -\log \int_{\mathbf{z}} \frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\mathrm{d}\mathbf{z} \\
&\leq -q_\phi(\mathbf{z}|\mathbf{x})\log \int_{\mathbf{z}} \frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\mathrm{d}\mathbf{z} \\
&= -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\log p_\theta(\mathbf{x}|\mathbf{z}) + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \qquad (8)
\end{aligned}$$

FVM

Factor VAE Metric (FVM) is a disentanglement metric that has a scale of 0 to 1, and the larger the better [11]. Let $(\mathbf{x}^{(1)}, ..., \mathbf{x}^{(L)})$ the random subset of training data with one ground-truth factor v_k is fixed. q_ϕ maps from $(\mathbf{x}^{(1)}, ..., \mathbf{x}^{(L)})$ to $(\mathbf{z}^{(1)}, ..., \mathbf{z}^{(L)})$, then let d is the dimension of $\mathbf{z}$ that has minimum variance. FVM indicates the accuracy of a majority-vote classifier that trains (d, k).

MIG

Mutual Information Gap (MIG) is a disentanglement metric that has a scale of 0 to 1, and the larger the better [2]. For the given ground-truth factor v_k, let $z_{i^{(k)}}$ and $z_{j^{(k)}}$ be the top two latent variables that have the highest mutual information with v_k. MIG is the average of mutual information gap, where K indicates the number of ground-truth factors.

$$MIG = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{H(v_k)}(I_n(z_{i^{(k)}}; v_k) - I_n(z_{j^{(k)}}; v_k)) \qquad (9)$$

A.5 Derivation of conditional denoising score matching

The score network $\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) : \mathbb{R}^{n+1+d} \to \mathbb{R}^n$ estimates score of conditional density function. Eq. (10) shows L2 loss of conditional score matching, but it is hard to train since $\nabla \text{log} p_t(\mathbf{x}_t|\mathbf{z})$ is intractable. This proof shows that optimize $\mathcal{L}_{cdsm}$ equals to optimize Eq. (10), since Eq. (12) shows the same form of $\mathcal{L}_{cdsm}$. Here, Eq. (11) is derived from Eq. (13).

$$\mathbb{E}_t \big[\lambda(t)\mathbb{E}_{p(\mathbf{x}_0)}[\|\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) - \nabla \text{log} p_t(\mathbf{x}_t|\mathbf{z})\|_2^2]\big] \tag{10}$$

$$= \mathbb{E}_t \big[\lambda(t)\mathbb{E}_{p(\mathbf{x}_0), p(\mathbf{x}_t|\mathbf{x}_0), q_\phi(\mathbf{z}|\mathbf{x}_0)}[\|\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) - \nabla \text{log} p_t(\mathbf{x}_t|\mathbf{z})\|_2^2]\big]$$

$$= \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t,\mathbf{z}}[\lambda(t)\ \|\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) - \nabla \text{log} p_t(\mathbf{x}_t|\mathbf{z})\|_2^2]]$$

$$= \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t,\mathbf{z}}[\lambda(t)\ \|\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) - \nabla \text{log} p_t(\mathbf{x}_t, \mathbf{z}) + \nabla \text{log} p(\mathbf{z})\|_2^2]]$$

$$= \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t,\mathbf{z}}[\lambda(t)\ \|\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) - \nabla \text{log} p_t(\mathbf{x}_t, \mathbf{z})\|_2^2]]$$

$$= \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t,\mathbf{z}}[\lambda(t)\ \|\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z})\|_2^2 - 2 \cdot \mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) \cdot \nabla \text{log} p_t(\mathbf{x}_t, \mathbf{z}) + c_1]$$

$$= \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t,\mathbf{z}}[\lambda(t)\ \|\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z})\|_2^2 - 2 \cdot \mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) \cdot \nabla \text{log} p_{0t}(\mathbf{x}_t|\mathbf{x}_0)] + c_1 \tag{11}$$

$$= \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t,\mathbf{z}}[\lambda(t)\ \|\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) - \nabla \text{log} p_{0t}(\mathbf{x}_t|\mathbf{x}_0)\|_2^2 - \|\nabla \text{log} p_{0t}(\mathbf{x}_t|\mathbf{x}_0)\|_2^2] + c_1$$

$$= \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t,\mathbf{z}}[\lambda(t)\ \|\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) - \nabla \text{log} p_{0t}(\mathbf{x}_t|\mathbf{x}_0)\|_2^2] - c_2 + c_1$$

$$= \mathbb{E}_t \big[\lambda(t)\mathbb{E}_{p(\mathbf{x}_0), p_{0t}(\mathbf{x}_t|\mathbf{x}_0), q_\phi(\mathbf{z}|\mathbf{x}_0)}[\|\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) - \nabla \text{log} p_{0t}(\mathbf{x}_t|\mathbf{x}_0)\|_2^2]\big] - c_2 + c_1 \tag{12}$$

$$\mathbb{E}_{\mathbf{x}_t,\mathbf{z}}[\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) \cdot \nabla \text{log} p_t(\mathbf{x}_t, \mathbf{z})] \tag{13}$$

$$= \int p_t(\mathbf{x}_t, \mathbf{z}) \cdot \mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) \cdot \nabla \text{log} p_t(\mathbf{x}_t, \mathbf{z}) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{z}$$

$$= \int \mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) \cdot \nabla p_t(\mathbf{x}_t, \mathbf{z}) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{z}$$

$$= \int \mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) \cdot \int \nabla p_t(\mathbf{x}_t, \mathbf{z}, \mathbf{x}_0) \mathrm{d}\mathbf{x}_0 \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{z}$$

$$= \int \mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) \cdot \int \nabla p_t(\mathbf{x}_t, \mathbf{z}|\mathbf{x}_0) p(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0 \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{z}$$

$$= \int \mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) \cdot p_t(\mathbf{x}_t, \mathbf{z}|\mathbf{x}_0) \cdot \int \nabla \text{log} p_t(\mathbf{x}_t, \mathbf{z}|\mathbf{x}_0) p(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0 \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{z}$$

$$= \int \mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) \cdot p_t(\mathbf{x}_t, \mathbf{z}|\mathbf{x}_0) \cdot p(\mathbf{x}_0) \cdot \int \nabla \text{log} p_t(\mathbf{x}_t, \mathbf{z}|\mathbf{x}_0) \mathrm{d}\mathbf{x}_0 \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{z}$$

$$= \mathbb{E}_{\mathbf{x}_0,\mathbf{x}_t,\mathbf{z}}[\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) \cdot \nabla \text{log} p_t(\mathbf{x}_t, \mathbf{z}|\mathbf{x}_0)]$$

$$= \mathbb{E}_{\mathbf{x}_0,\mathbf{x}_t,\mathbf{z}}\big[\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) \cdot [\nabla \text{log} p_{0t}(\mathbf{x}_t|\mathbf{x}_0) + \nabla \text{log} p(\mathbf{z}|\mathbf{x}_0)]\big]$$

$$= \mathbb{E}_{\mathbf{x}_0,\mathbf{x}_t,\mathbf{z}}[\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{z}) \cdot \nabla \text{log} p_{0t}(\mathbf{x}_t|\mathbf{x}_0)]$$