Why Are Conditional Generative Models Better Than Unconditional Ones?

Anonymous Author(s)
Affiliation
Address
email

Abstract

Extensive empirical evidence demonstrates that conditional generative models are easier to train and perform better than unconditional ones by exploiting the labels of data. So do score-based diffusion models. In this paper, we analyze the phenomenon formally and identify that the key of conditional learning is to partition the data properly. Inspired by the analyses, we propose *self-conditioned diffusion models* (SCDM), which is trained conditioned on indices clustered by the k-means algorithm on the features extracted by a model pre-trained in a self-supervised manner. SCDM significantly improves the unconditional model across various datasets and achieves a record-breaking FID of 3.94 on ImageNet 64x64 without labels. Besides, SCDM achieves a slightly better FID than the corresponding conditional model on CIFAR10.

1 Introduction

Extensive empirical evidence in prior work [14, 3, 9] demonstrates that conditional generative models are easier to train and perform better than unconditional ones by exploiting the labels of data. So do score-based diffusion models (DM). For instance, the representative work [9] achieves a FID of 10.94 when trained conditionally and a FID of 26.21 when trained unconditionally on ImageNet of size 256x256.

Intuitively, the gap exists because (1) the marginal distribution induced by a conditional model is more expressive than the corresponding unconditional model; and (2) the data distribution conditioned on a specific class has fewer modes and is easier to fit than the original data distribution.

In this paper, we formalize the above intuition in an ideal setting where we have infinite data. It is easy to show that the marginal distribution induced by a conditional model can be viewed as a mixture of the corresponding unconditional models. Further, we derive a sufficient condition for the superiority of the conditional model, which suggests that the conditional model gains more as the conditional data distribution gets simpler. The analyses explain previous empirical findings: conditioning on class labels probably partitions the data into simpler groups according to the semantics of data.

Notably, our analyses apply to all possible conditions, not limited to class labels. Then, a very natural idea is to find a certain way to obtain meaningful conditions in an unsupervised manner and boost the unconditional generation results. The recent advances in self-supervised learning [10, 5] show that one can learn predictive representations without labels, which serve as an ideal tool for obtaining meaningful conditions. Specifically, we simply run a clustering algorithm (e.g., k-means) on the features extracted by a model pre-trained in a self-supervised manner (on the same dataset) and use the cluster indices as conditions to train a conditional model.

Although our analyses and the self-conditional approach is applicable to all types of deep generative models, we focus on score-based diffusion models in our experiments to explore the boundary of

Under review at the NeurIPS 2022 Workshop on Score-Based Methods. Do not distribute.

unsupervised generative modeling. Therefore, we refer to our approach as *self-conditioned diffusion models* (SCDM). We systematically evaluate SCDM on several widely adopted datasets. In all settings, SCDM significantly improves the unconditional model. Notably, SCDM achieves a record-breaking FID of 3.94 on ImageNet 64x64 without labels. Besides, SCDM achieves a slightly better FID than the corresponding conditional model on CIFAR10.

2 Why Are Conditional Generative Models Better Than Unconditional Ones

In this section, we present the problem formulation and our analyses.

2.1 Problem Formulation

Let $q(\boldsymbol{x}, c)$ be the joint distribution of the data $\boldsymbol{x}$ and the condition c and $q(\boldsymbol{x}) := \sum_c q(\boldsymbol{x}, c)$. Let $p_{\boldsymbol{\theta},E}(\boldsymbol{x})$ be a model parameterized by $\boldsymbol{\theta} \in \Theta$ and $E \in \mathcal{E}$, where $\boldsymbol{\theta}$ denotes the parameters in the backbone and E is the embedding for a condition. We formalize two learning paradigms as follows.

In *unconditional learning*, $p_{\boldsymbol{\theta},E}(\boldsymbol{x})$ approximates the marginal data distribution $q(\boldsymbol{x})$ directly and E is a redundant embedding shared by all data. Formally, given a certain statistics divergence $\mathcal{D}$ (or more loosely a divergence upper bound [20, 2]), unconditional learning aims to optimize

$$\min_{\boldsymbol{\theta}\in\Theta, E\in\mathcal{E}} \mathcal{D}(q(\boldsymbol{x}) \| p_{\boldsymbol{\theta},E}(\boldsymbol{x})). \tag{1}$$

In *conditional learning*, the embedding E is spared to receive the signal from the condition c, through an embedding function $\boldsymbol{\phi} \in \Phi$. This induces a conditional model $p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}|c) := p_{\boldsymbol{\theta},E}(\boldsymbol{x})|_{E=\boldsymbol{\phi}(c)}$, which approximates the conditional data distribution $q(\boldsymbol{x}|c)$ by tuning the backbone $\boldsymbol{\theta}$ and the embedding function $\boldsymbol{\phi}$. Formally, conditional learning aims to optimize

$$\min_{\boldsymbol{\theta}\in\Theta, \boldsymbol{\phi}\in\Phi} \mathbb{E}_{q(c)} \mathcal{D}(q(\boldsymbol{x}|c) \| p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}|c)). \tag{2}$$

The conditional model applies ancestral sampling to generate samples, where a condition c is firstly drawn from $q(c)$[1], and then a data $\boldsymbol{x}$ is drawn from $p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}|c)$. Such a process produces samples from $p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}) := \mathbb{E}_{q(c)} p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}|c)$. The generation performance of the conditional model is evaluated according to how close $p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x})$ is to the data distribution $q(\boldsymbol{x})$, i.e., $\mathcal{D}(q(\boldsymbol{x}) \| p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}))$.

2.2 Analyses

In this section, we attempt to formalize two insights on why conditional learning of generative models generally outperforms the unconditional one.

Firstly, we compare the expressive power of the two strategies with the same backbone parameterized by $\boldsymbol{\theta}$. As shown in Section 2.1, the conditional model produces samples from $p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}) = \mathbb{E}_{q(c)}[p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}|c)] = \mathbb{E}_{q(c)}[p_{\boldsymbol{\theta},E}(\boldsymbol{x})|_{E=\boldsymbol{\phi}(c)}]$. Therefore, $p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x})$ can be viewed as a mixture of several unconditional models. Namely, the conditional model is more expressive than the unconditional one, despite the fact that both models are based on the same backbone $p_{\boldsymbol{\theta},E}(\boldsymbol{x})$.

Secondly, we derive a sufficient condition for the superiority of the conditional model. Let $\boldsymbol{\theta}_u^*, E_u^*$ be the optimal solution of the unconditional learning in Eq. (1). Let $\boldsymbol{\theta}_c^*, \boldsymbol{\phi}_c^*$ be the optimal solution of the conditional learning in Eq. (2). Proposition 1 characterizes a sufficient condition for $\mathcal{D}(q(\boldsymbol{x}) \| p_{\boldsymbol{\theta}_c^*,\boldsymbol{\phi}_c^*}(\boldsymbol{x})) < \mathcal{D}(q(\boldsymbol{x}) \| p_{\boldsymbol{\theta}_u^*,E_u^*}(\boldsymbol{x}))$.

Proposition 1. *Suppose for any parameter $\boldsymbol{\theta} \in \Theta$ and any condition c, approximating $q(\boldsymbol{x}|c)$ is simpler than $q(\boldsymbol{x})$ by only tuning the embedding E of $p_{\boldsymbol{\theta},E}(\boldsymbol{x})$, i.e., $\min_E \mathcal{D}(q(\boldsymbol{x}|c) \| p_{\boldsymbol{\theta},E}(\boldsymbol{x})) < \min_E \mathcal{D}(q(\boldsymbol{x}) \| p_{\boldsymbol{\theta},E}(\boldsymbol{x}))$. Then, under additional mild regularity conditions [2], $\mathcal{D}(q(\boldsymbol{x}) \| p_{\boldsymbol{\theta}_c^*,\boldsymbol{\phi}_c^*}(\boldsymbol{x})) < \mathcal{D}(q(\boldsymbol{x}) \| p_{\boldsymbol{\theta}_u^*,E_u^*}(\boldsymbol{x}))$ holds. (Proof in Appendix A)*

[1] We assume $q(c)$ is known, which is satisfied in conditional learning with labels.

[2] Specifically, we assume that the divergence $\mathcal{D}$ is convex and the embedding function space Φ includes all measurable functions, which are verifiable in practice. In fact, the former can be satisfied using the KL divergence and the latter can be satisfied by using nonparametric embeddings.

Table 1: FID↓ results on different datasets. K represents the number of clusters.

	CIFAR10	CelebA 64x64	LSUN Bedroom 64x64	ImageNet 64x64
Unconditional DM	2.72	2.14	2.69	6.44
Conditional DM	2.24	-	-	**3.08**
SCDM ($K = 2$)	-	2.04	-	-
SCDM ($K = 10$)	**2.23**	**1.91**	-	-
SCDM ($K = 20$)	2.27	2.08	2.39	-
SCDM ($K = 30$)	2.30	-	-	-
SCDM ($K = 50$)	2.34	-	-	-
SCDM ($K = 100$)	-	-	**2.25**	-
SCDM ($K = 1000$)	-	-	-	3.94

The sufficient condition in Proposition 1 is hard to verify in practice generally[3]. However, it does provide insights on when conditional learning is preferable. In fact, it implies that the conditional model gains more (i.e., $\min_E \mathcal{D}(q(\boldsymbol{x}|c)\|p_{\boldsymbol{\theta},E}(\boldsymbol{x}))$ gets smaller for all $\boldsymbol{\theta}$) as the conditional data distribution gets simpler. The condition is probably satisfied in practical conditional learning with class labels. In this sense, Proposition 1 explains previous empirical findings.

3 Self-Conditioned Diffusion Models

Note that Proposition 1 applies to all possible conditions, not limited to class labels, which inspires us to obtain meaningful conditions in an unsupervised manner to boost the unconditional generation results. The recent advances in self-supervised learning [10, 5] show that one can learn predictive representations without labels, which serves as an ideal tool for obtaining meaningful conditions.

Specifically, we propose a three-stage algorithm. Firstly, we train a feature extractor on the target dataset (without labels) in a self-supervised manner and extract features. Secondly, we run a clustering algorithm (e.g., k-means in our experiments) on these features and obtain the cluster indices for all data. Finally, we train a conditional diffusion model [18, 9] by taking the cluster indices as conditions. We refer to our approach as *self-conditioned diffusion models* (SCDM).

We mention that the high-level idea of using clustering indices from self-supervised learning coincides with prior work in GANs [1, 4, 19]. This paper presents distinct contributions in the following aspects. First, prior work focuses on avoiding mode collapse while this paper is motivated by a different perspective with theoretical insights missing in the literature. Second, this paper is built upon SOTA diffusion models [6, 9] to explore the boundary of unconditional generative modeling. In fact, we obtain a record-breaking FID of 3.94 on ImageNet 64x64 without labels. See a direct comparison with prior work [4, 19] in Table 2.

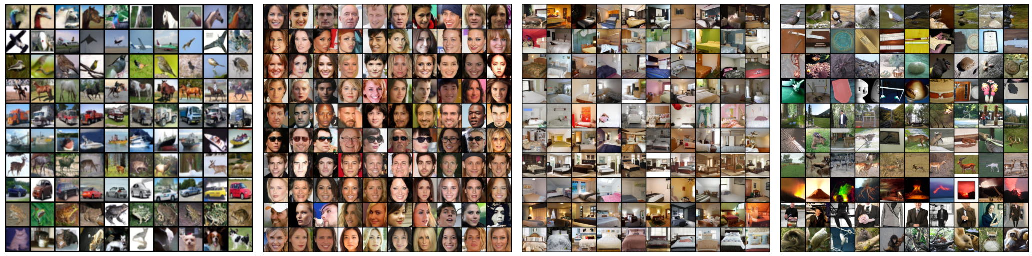
4 Experiment

We evaluate SCDM on CIFAR10 [13], CelebA 64x64 [15], LSUN Bedroom 64x64 [22] and ImageNet 64x64 [8]. By default, we use MoCo-v2 [6] on CIFAR10, CelebA 64x64 and LSUN Bedroom 64x64, and use MoCo-v3 [7] on ImageNet 64x64, in the self-supervised learning stage. We use the FID score [11] to measure the sample quality. We use the same architecture for SCDM and its unconditional and conditional baselines. See more experimental details in Appendix B.

4.1 Sample Quality

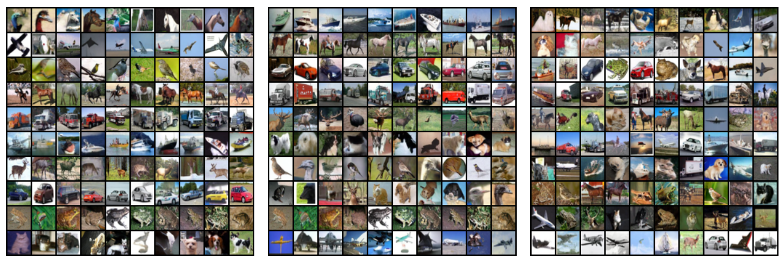
Firstly, we compare our SCDM with the unconditional and conditional baselines. As shown in Table 1, SCDM uniformly outperforms the unconditional model and slightly outperforms the conditional

[3]A simple verifiable case is to fit a mixture of Gaussian (MoG) data by a single Gaussian (unconditional learning) or a MoG with ground-truth cluster indices (conditional learning).



(a) CIFAR10 (b) CelebA 64x64 (c) LSUN Bedroom 64x64 (d) ImageNet 64x64

Figure 2: Generated samples of SCDM. Each column corresponds to a cluster. We use the model with the best FID.



(a) MoCo-v2 ($K = 10$) (b) SimCLR ($K = 10$) (c) Pixel ($K = 10$)

Figure 3: Generated samples on CIFAR10 with different clustering methods.

model on CIFAR10. On ImageNet 64x64, SCDM greatly improves the FID compared to the unconditional model. We provide generated samples in Figure 2.

In Table 2, we compare SCDM with other methods on ImageNet 64x64 in the unlabelled setting. SCDM significantly outperforms all prior methods and achieves a record-breaking FID of 3.94.

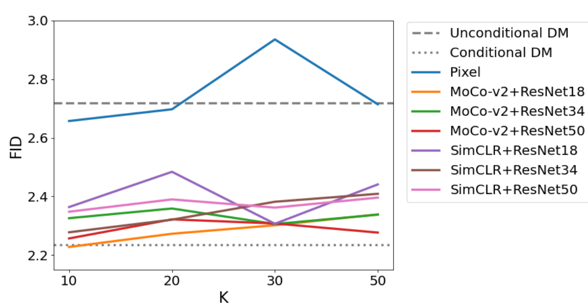


Figure 1: The effect of the self-supervised learning methods, and the backbones used in self-supervised learning.

Table 2: ImageNet 64x64 results in the unlabelled setting. †Improved DDPM reports FID with 10K samples, and thereby we use reproduced results on 50K samples [2].

Method	FID
SLCGAN [19]	19.2
Unconditional BigGAN [4]	16.9
IC-GAN [4]	9.2
Improved DDPM† [18]	16.38
Unconditional DM	6.44
SCDM (ours)	**3.94**

4.2 Ablation Study

In this part, we study the effect of the self-supervised learning methods. We test MoCo-v2, as well as SimCLR [5] with 3 backbones: ResNet-18, ResNet-34, and ResNet-50. We also perform k-means on image pixels directly to get cluster indices, and we call this method *pixel*. As shown in Figure 1, SimCLR performs similarly to MoCo-v2, and the choice of backbones does not affect the performance much. However, k-means on image pixels performs much worse than SimCLR and MoCo-v2. Indeed, as shown in Figure 3, we find objects of diverse classes appear in a single cluster for the pixel method, leading to a more complex distribution in a single cluster, which is more difficult to learn.

References

[1] Mohammadreza Armandpour, Ali Sadeghian, Chunyuan Li, and Mingyuan Zhou. Partition-guided gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5099–5109, 2021.

[2] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[7] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[11] 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.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[14] Chongxuan Li, Taufik Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. *Advances in neural information processing systems*, 30, 2017.

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3730–3738. IEEE Computer Society, 2015.

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[17] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022.

[18] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021.

[19] Mehdi Noroozi. Self-labeled conditional gans. *arXiv preprint arXiv:2012.02162*, 2020.

[20] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *arXiv e-prints*, pages arXiv–2101, 2021.

[21] 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.

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

A Proof of Proposition 1

We firstly present a lemma.

Lemma 1. *Suppose* $\min_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{q(c)} \min_{E\in\mathcal{E}} \mathcal{D}(q(\boldsymbol{x}|c)\|p_{\boldsymbol{\theta},E}(\boldsymbol{x})) < \min_{\boldsymbol{\theta}\in\Theta, E\in\mathcal{E}} \mathcal{D}(q(\boldsymbol{x})\|p_{\boldsymbol{\theta},E}(\boldsymbol{x}))$*, the divergence* $\mathcal{D}$ *is convex, and the embedding function space* Φ *includes all measurable functions. Then we have* $\mathcal{D}(q(\boldsymbol{x})\|p_{\boldsymbol{\theta}_c^*,\boldsymbol{\phi}_c^*}(\boldsymbol{x})) < \mathcal{D}(q(\boldsymbol{x})\|p_{\boldsymbol{\theta}_u^*,E_u^*}(\boldsymbol{x}))$.

Proof. According to the convexity of $\mathcal{D}$, we have

$$\mathcal{D}(q(\boldsymbol{x})\|p_{\boldsymbol{\theta}_c^*,\boldsymbol{\phi}_c^*}(\boldsymbol{x})) = \mathcal{D}(\mathbb{E}_{q(c)}q(\boldsymbol{x}|c)\|\mathbb{E}_{q(c)}p_{\boldsymbol{\theta}_c^*,\boldsymbol{\phi}_c^*}(\boldsymbol{x}|c)) \leq \mathbb{E}_{q(c)}\mathcal{D}(q(\boldsymbol{x}|c)\|p_{\boldsymbol{\theta}_c^*,\boldsymbol{\phi}_c^*}(\boldsymbol{x}|c)) \quad (3)$$

According to the definition of $\boldsymbol{\theta}_c^*, \boldsymbol{\phi}_c^*$, we have

$$\begin{aligned}&\mathbb{E}_{q(c)}\mathcal{D}(q(\boldsymbol{x}|c)\|p_{\boldsymbol{\theta}_c^*,\boldsymbol{\phi}_c^*}(\boldsymbol{x}|c)) = \min_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}_{q(c)}\mathcal{D}(q(\boldsymbol{x}|c)\|p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}|c))\\ =&\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}_{q(c)}\mathcal{D}(q(\boldsymbol{x}|c)\|p_{\boldsymbol{\theta},\boldsymbol{\phi}(c)}(\boldsymbol{x})) = \min_{\boldsymbol{\theta}} \mathbb{E}_{q(c)} \min_{\boldsymbol{\phi}(c)} \mathcal{D}(q(\boldsymbol{x}|c)\|p_{\boldsymbol{\theta},\boldsymbol{\phi}(c)}(\boldsymbol{x}))\\ =&\min_{\boldsymbol{\theta}} \mathbb{E}_{q(c)} \min_{E} \mathcal{D}(q(\boldsymbol{x}|c)\|p_{\boldsymbol{\theta},E}(\boldsymbol{x})). \end{aligned} \quad (4)$$

Combining Eq. (3), Eq. (4), and the assumption, we have

$$\begin{aligned}&\mathcal{D}(q(\boldsymbol{x})\|p_{\boldsymbol{\theta}_c^*,\boldsymbol{\phi}_c^*}(\boldsymbol{x})) \leq \min_{\boldsymbol{\theta}} \mathbb{E}_{q(c)} \min_{E} \mathcal{D}(q(\boldsymbol{x}|c)\|p_{\boldsymbol{\theta},E}(\boldsymbol{x}))\\ <&\min_{\boldsymbol{\theta},E} \mathcal{D}(q(\boldsymbol{x})\|p_{\boldsymbol{\theta},E}(\boldsymbol{x})) = \mathcal{D}(q(\boldsymbol{x})\|p_{\boldsymbol{\theta}_u^*,E_u^*}(\boldsymbol{x})).\end{aligned}$$

□

Then we present proof of Proposition 1.

Proof. Since $\forall \boldsymbol{\theta}, c, \min_{E} \mathcal{D}(q(\boldsymbol{x}|c)\|p_{\boldsymbol{\theta},E}(\boldsymbol{x})) < \min_{E} \mathcal{D}(q(\boldsymbol{x})\|p_{\boldsymbol{\theta},E}(\boldsymbol{x}))$, we have

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{q(c)} \min_{E} \mathcal{D}(q(\boldsymbol{x}|c)\|p_{\boldsymbol{\theta},E}(\boldsymbol{x})) < \min_{\boldsymbol{\theta},E} \mathcal{D}(q(\boldsymbol{x})\|p_{\boldsymbol{\theta},E}(\boldsymbol{x})).$$

According to Lemma 1, we have $\mathcal{D}(q(\boldsymbol{x})\|p_{\boldsymbol{\theta}_c^*,\boldsymbol{\phi}_c^*}(\boldsymbol{x})) < \mathcal{D}(q(\boldsymbol{x})\|p_{\boldsymbol{\theta}_u^*,E_u^*}(\boldsymbol{x}))$. □

B Experimental Details

In the self-supervised learning stage, we use ResNet18 on CIFAR10, and ResNet50 on CelebA and LSUN Bedroom. We train 1600, 800 and 200 epochs on CIFAR10, CelebA and LSUN Bedroom respectively. We resize images to 32x32, and use a batch size of 512. We use the SGD optimizer with a learning rate of 0.06, a momentum of 0.9 and a weight decay of 5e-4. The queue size of MoCo is 12800, and the momentum of MoCo is 0.999. As for ImageNet, we use pretrained ViT-Base provided in `https://github.com/facebookresearch/moco-v3`.

We provide training and sampling setting of diffusion models in Table 3. We evaluate FID every 50K training steps, and report the best one.

Dataset	CIFAR10	CelebA 64x64	LSUN Bedroom 64x64	ImageNet 64x64
Architecture	IDDPM [18]	IDDPM	IDDPM	ADM [9]
Noise schedule	VP SDE [21]	VP SDE	VP SDE	VP SDE
Batch size	128	128	128	2048
Training steps	1M	1M	1M	550K
Optimizer	Adam [12]	Adam	Adam	AdamW [16]
Learning rate	1e-4	1e-4	1e-4	3e-4
Sampler	EM	EM	EM	DPM-Solver [17]
Sampling steps	1K	1K	1K	50

Table 3: The experimental setting of diffusion models. EM represents the Euler-Maruyama sampler.